J Am Acad Audiol 29:164–174 (2018)

Test–Retest Reliability of Dual-Recorded Brainstem versus Cortical Auditory-Evoked Potentials to Speech

DOI: 10.3766/jaaa.16167

Gavin M. Bidelman*†‡
Monique Pousson†
Calli Dugas†
Amy Fehrenbach†

Abstract

Background: Auditory-evoked potentials have proven useful in the objective evaluation of sound encoding at different stages of the auditory pathway (brainstem and cortex). Yet, their utility for use in clinical assessment and empirical research relies critically on the precision and test–retest repeatability of the measure.

Purpose: To determine how subcortical/cortical classes of auditory neural responses directly compare in terms of their internal consistency and test–retest reliability within and between listeners.

Research Design: A descriptive cohort study describing the dispersion of electrophysiological measures.

Study Sample: Eight young, normal-hearing female listeners.

Data Collection and Analysis: We recorded auditory brainstem responses (ABRs), brainstem frequency-following responses (FFRs), and cortical (P1-N1-P2) auditory-evoked potentials elicited by speech sounds in the same set of listeners. We reassessed responses within each of four different test sessions over a period of 1 mo, allowing us to detect possible changes in latency/amplitude characteristics with finer detail than in previous studies.

Results: Our findings show that brainstem and cortical amplitude/latency measures are remarkably stable; with the exception of slight prolongation of the P1 wave, we found no significant variation in any response measure. Intraclass correlation analysis revealed that the speech-evoked FFR amplitude and latency measures achieved superior repeatability (intraclass correlation coefficient >0.85) among the more widely used obligatory brainstem (ABR) and cortical (P1-N1-P2) auditory-evoked potentials. Contrasting these intersubject effects, intrasubject variability (i.e., within-subject coefficient of variation) revealed that while latencies were more stable than amplitudes, brainstem and cortical responses did not differ in their variability at the single subject level.

Conclusions: We conclude that (1) the variability of auditory neural responses increases with ascending level along the auditory neuroaxis (cortex $>$ brainstem) between subjects but remains highly stable within subjects and (2) speech-FFRs might provide a more stable measure of auditory function than other conventional responses (e.g., click-ABR), given their lower inter- and intrasubject variability.

Key Words: auditory brainstem response, event-related brain potentials, frequency-following response, intraclass correlation coefficient, test–retest reliability

Abbreviations: ABR = auditory brainstem response; AEP = auditory-evoked potential; CV = coefficient of variation; EEG = electroencephalogram; ERP = event-related brain potentials; F0 = fundamental frequency; FFR = frequency-following responses; ICC = intraclass correlation coefficient; SNR = signal-to-noise ratio

*Institute for Intelligent Systems, University of Memphis, Memphis, TN; †School of Communication Sciences and Disorders, University of Memphis, Memphis, TN; ‡Department of Anatomy and Neurobiology, University of Tennessee Health Sciences Center, Memphis, TN

Corresponding author: Gavin M. Bidelman, School of Communication Sciences and Disorders, University of Memphis, Memphis, TN 38152; E-mail: g.bidelman@memphis.edu

This work was supported by grants from the American Hearing Research Foundation (AHRF) and American Academy of Audiology (AAA) Foundation awarded to G.M.B.

INTRODUCTION

Auditory-evoked potential (AEP) measures such as the auditory brainstem response (ABR) are used routinely to noninvasively assess the functional integrity of auditory structures, threshold estimation, and differential diagnosis in suspected cases of retrocochlear pathology in both adults and infants (Sininger et al, 1997; Stapells, 2000). These electrophysiological measures are also used in assessment batteries for central auditory processing disorders (Jerger and Musiek, 2000) and auditory neuropathy (Starr et al, 1996). An important issue in AEP testing is the reliability (i.e., test–retest consistency) of neural responses. Low variability within and between listeners ensures reliable and consistent results in both research and clinical investigation.

In this regard, the auditory cortical event-related brain potential (ERP) components (P1-N1-P2) have often been dismissed in objective audiometry, given their later maturational time course (and thus inappropriateness for testing young children) (Ponton et al, 2000), dependence on subject arousal and attention (e.g., Picton et al, 1971; Näätänen, 1992), and presumed higher response variability than the early components (Jerger and Jerger, 1985; Hall, 1992). However, empirical studies reveal that the cortical ERPs can actually be more accurate than traditional ABRs in estimating thresholds, with 94% of estimates within 10–15 dB of behavior (e.g., Lightfoot and Kennedy, 2006). Consequently, understanding the inter- and intrasubject variability in the cortical ERPs would be important for establishing their clinical utility.

Repeatability of the cortical potentials has been assessed over periods ranging from several weeks to years (Walhovd and Fjell, 2002; Tremblay et al, 2003; Tervaniemi et al, 2005; Williams et al, 2005; McFadden et al, 2014). In their analysis of test–retest reliability of speech-evoked ERPs, Tremblay et al (2003) reported an average intraclass correlation coefficient (ICC) of 0.86 ± 0.08 across 28 different stimulus conditions tested over the period of ~1 week (see their Figure 3). These findings demonstrate high repeatability of the cortical ERPs. However, this study only considered consistency of gross response morphology (i.e., ICCs were computed on whole waveforms), rather than replication of individual response metrics (e.g., peak latency and amplitudes). Similar repeatability has been reported for the auditory mismatch negativity, which shows intrasession correlations ranging from $r = 0.49$–0.89 when tested within ~8 days (Tervaniemi et al, 2005) to $r \approx 0.80$ for test–retest intervals up to 12–16 weeks (Salinsky et al, 1991).

Response reliability has also been assessed for AEPs emitted from subcortical auditory structures. ABRs—routine in audiological testing—show remarkably stable latency characteristics (Edwards et al, 1982; Oyler et al, 1991), such that submillisecond variation in neural responses can be used for diagnostic purposes (e.g., detecting hearing loss; Hall, 1992; Picton, 2010). Fewer studies have assessed reliability for the brainstem frequency-following response (FFR). FFR is a "neurophonic" potential generated in the upper brainstem (Bidelman, 2015b) that closely mirrors spectrotemporal properties of acoustic stimuli (Krishnan, 2007; Skoe and Kraus, 2010; Bidelman, 2015b). FFRs provide a detailed window into the encoding of complex sounds within the human electroencephalogram (EEG) as evident by the fact they are actually intelligible to human listeners when replayed as auditory stimuli (Galbraith et al, 1995; Weiss and Bidelman, 2015). FFRs are functionally distinct from ABRs (Song et al, 2006; Bidelman, 2015b). Although not yet in the clinical mainstream, brainstem FFRs may offer a unique window into normal and disordered auditory brain processing not afforded by other electrophysiological measures. For instance, speech-evoked FFRs are impaired in children with language and learning disorders (Banai et al, 2007; Banai et al, 2009; Basu et al, 2010; Rocha-Muniz et al, 2012) despite normal click-ABRs (Song et al, 2006). These findings imply that the FFR might offer important diagnostic function of real-world speech listening skills not tapped by conventional audiological evaluation (for reviews, see Chandrasekaran and Kraus, 2010; Bidelman, 2017; Kraus et al, 2017).

Of the few studies examining FFR reliability, Song et al (2011) measured responses over the period of ~2 mo between test sessions. Analysis of responses was assessed in both the frequency and time domains. This study found several response measures including peak latencies, stimulus-to-response consistency, and spectral amplitudes to be highly replicable between tests. However, the strength of test–retest correlations varied considerably depending on the specific response metric ($r_{spearman} = 0.12$–0.82), leading some to debate the clinical utility of the FFR (e.g., McFarland and Cacace, 2012). In the present study, we aimed to further assess test–retest reliability of the speech-FFR using a more comprehensive testing schedule (×4 sessions) than available in previous studies as well as document both within (intra-) and between (inter-) subject variability in this measure.

Another issue in directly comparing AEP repeatability is the inherent difference in variability and measurement scale between different classes of response (i.e., brainstem versus cortical potentials). Cortical ERPs, for example, are thought to be more labile than the brainstem potentials due to their susceptibility to subject factors (i.e., attention) and stronger adaptation of cortical compared to subcortical generators (Thornton and Coleman, 1975; Picton et al, 1978; Dalebout and Robey, 1997; Bidelman, 2015a). Indeed, previous

reliability studies have concluded that brainstem components (e.g., ABR wave V) are generally less variable at the individual subject level than latter components (Lauter and Karzon, 1990a,b). Moreover, millisecond delays in brainstem potentials can be diagnostically meaningful (e.g., signal a hearing loss) (Hall, 1992; Picton, 2010), whereas similar prolongations in the cortical ERPs might be considered normal response variation. To this end, our second aim was to directly evaluate how brainstem (ABR and FFR) and cortical (ERP) classes of the AEP directly compare in terms of their internal consistency and test–retest reliability.

The current study addresses subject test–retest reliability across a thorough testing regimen, and additionally compares the intra/inter subject variability of cortical versus brainstem potentials. Our design included four recordings distributed evenly across the course of 1 mo. This allowed us to determine whether the AEPs show subtle variation (e.g., in amplitude or latency characteristics) with a finer detail than previous studies employing only one test–retest evaluation (e.g., Tremblay et al, 2003; Tervaniemi et al, 2005; Song et al, 2011). Our second aim was to directly compare reliability of the different classes of evoked potentials. We measured both speech-evoked brainstem FFRs, conventional click-evoked ABRs, and the obligatory cortical ERPs (P1-N1-P2). This allowed us to compare test–retest reliability across subcortical and cortical AEP classes within the same set of listeners. Our findings reveal remarkably stable responses across both brainstem and cortical potentials with superior reliability for the brainstem FFR.

METHODS

Participants

Eight young adults (age range: 18–35 yr, all females) participated in the experiment. This sample size was determined adequate based on our previous work (Bidelman, 2015a) using comparable numbers and the fact that stimulus-related differences emerged in P1 and N1 with large effect sizes (Cohen's $d > 1.2$) and >88% power (Power and Precision™; Bornstein et al, 1997). All participants were native speakers of English, right handed, had normal hearing (i.e., audiometric thresholds ≤25 dB HL; 500–4000 Hz), and reported no previous history of neuropsychiatric illnesses. Musical training is known to enhance brainstem and cortical AEPs (e.g., Bidelman, Weiss, et al, 2014; Bidelman and Alain, 2015; Weiss and Bidelman, 2015). Hence, all participants were required to have minimal (<5 yr) musical training. Participants were paid and gave written informed consent in compliance with a protocol approved by the Institutional Review Board of the University of Memphis.

Test Session Schedule

The timing of recording sessions followed a rigorous and highly controlled testing sequence. Participants returned for electrophysiological testing in each of four sessions (spaced ~1 week apart) across the course of a month. Each session occurred in the morning between 9:30 and 11:00 AM to limit variations in circadian rhythm effects (e.g., body temperature) across test sessions. In addition, the testing schedule was initiated according to the start of each participant's menstrual cycle—corroborated via an over-the-counter ovulation test (Clearblue®). Subsequent sessions occurred every 7 ± 2 days thereafter, so as to align with each quartile of the menstrual cycle. This synchronization of testing helped further control intersubject variation in the AEPs, which can arise due to hormone-induced changes in neural inhibition (Elkind-Hirsch et al, 1992; Caruso et al, 2003).

Electrophysiological Recordings

All electrophysiological testing was administered by a clinical audiologist (MP), which ensured identical electrode placements and recording conditions for each session. For each recording, participants reclined comfortably in an electro-acoustically shielded booth to facilitate recording of neurophysiological responses. Attention is known to have a differential effect between subcortical and cortical levels of the auditory system, having a stronger effect on the cortical ERPs (Picton and Hillyard, 1974; Woods and Hillyard, 1978; Hillyard and Picton, 1979; Okamoto et al, 2011). They were instructed to relax and refrain from extraneous body movement and ignore the sounds they hear (to divert attention away from the stimuli), and were allowed to watch a muted subtitled movie to maintain a calm yet wakeful state. This allowed us to record each type of AEP in a passive listening paradigm while controlling arousal and attentional state. Stimulus presentation was controlled by MATLAB® 2014 (The MathWorks) routed to a TDT RP2 interface (Tucker-Davis Technologies, Alachua, FL). All stimuli were delivered binaurally at an intensity of 80 dB SPL through shielded insert earphones (ER-2, Etymotic Research, Elk Grove Village, IL).

Dual Brainstem-Cortical Speech-Evoked Potentials (FFRs/ERPs)

We used a 100-msec synthetic vowel (/a/) constructed using a cascade formant synthesizer (Klatt and Klatt, 1990) to elicit subcortical and cortical speech-ERPs (for details, see Bidelman et al, 2013). The speech sound was characterized by steady-state fundamental (F0) and formant (F1–F4) frequencies (F0: 100, F1: 730, F2: 1090, and F3: 2350 Hz). Using this speech token,

brainstem FFRs were recorded concurrently with the cortical ERPs in an optimized paradigm (for details, see Bidelman, 2015a). Briefly, this stimulus sequencing used two different interstimulus intervals and number of trials to optimally evoke the brainstem FFR and cortical ERP. This stimulus sequencing was advantageous as it allowed the collection of both response classes quasi-simultaneously, with minimal response habituation, and in roughly one-third the time of conventional (fixed interstimulus interval) presentation (Bidelman, 2015a). Following our previous report establishing this method (Bidelman, 2015a), two cortical responses were recorded for every 14 brainstem responses (FFR/ERP ratio = 14:2). In the current study, the total runtime included 3,500 sweeps for constructing the final brainstem FFR average and 500 trials for the cortical ERP.

The larger number of trials necessary for FFRs compared to ERPs is due to the fact that cortical responses are roughly an order of magnitude larger than brainstem responses (i.e., ERP [μV] versus FFR [nV] range). This means that ERPs require fewer sweeps to detect and achieve adequate signal-to-noise ratio (SNR) (Chandrasekaran and Kraus, 2010; Bidelman, 2014; 2015a). Nevertheless, the residual noise in the AEPs improves with $\sqrt{N}$, so different trial numbers between AEP classes might render comparison between levels spurious due to simple differences in signal quality. To rule out this possibility, we measured the SNR of each AEP response per session. SNR was computed as the ratio of signal amplitude to the standard deviation within the poststimulus epoch window (i.e., SNR = AEP_{amp}/σ_{epoch}), where σ_{epoch} is an estimate of noise overlapping with the evoked AEP (Hu et al, 2010). AEP_{amp} was taken as the wave V amplitude, F0 amplitude, and N1 amplitude for the ABR, FFR, and ERP, respectively, as these were the most prominent amplitude signatures of each response. Critically, SNR did not differ across test session [$F_{(3,77)} = 0.39$, $p = 0.76$] nor AEP class [$F_{(2,77)} = 2.65$, $p = 0.08$], indicating that brainstem/cortical responses were not inherently noisier than one another or between days.

Neuroelectric activity was recorded differentially between Ag/AgCl disc electrodes placed on the scalp at the high forehead at the hairline (∼Fpz) referenced to linked mastoids (A1/A2). Another electrode placed on the midforehead served as common ground. This vertical electrode montage is optimal for the simultaneous recording of brainstem and cortical AEPs (Musacchia et al, 2008; Bidelman et al, 2013; Bidelman, 2015a,b). Interelectrode impedance was maintained ≤3 kΩ for all recordings. Continuous EEGs were digitized at 10 kHz (SynAmps RT amplifiers; Compumedics Neuroscan, Charlotte, NC) using an online filter passband of DC to 4000 Hz. This high sampling rate was necessary to digitize the fast, phase-locked components of the brainstem FFR. Traces were then segmented (cortical ERP: −100 to 500 msec; brainstem FFR: 0 to 140 msec), baselined to the prestimulus interval, and subsequently averaged in the time domain to obtain evoked responses for each condition. Trials less than ±50 μV were rejected as artifacts prior to averaging. Evoked responses were then bandpass filtered into high-frequency (90–2000 Hz) and low-frequency (3–40 Hz) bands to isolate brainstem and cortical activity, respectively (e.g., Musacchia et al, 2008; Bidelman et al, 2013).

ABR

Click-evoked ABRs were recorded using an identical electrode configuration (Fpz-A1/A2) and sample rate (10 kHz) as the FFR and ERP recordings. This ensured that the precision of digitization rate was identical between AEP classes, allowing us to directly compare variability and rule out differences in scale of measurement that can arise using conventional (different) sample rates between response types (e.g., 500 Hz for cortical ERPs and 10 kHz for ABR/FFR). ABRs were evoked in response to a 100-μsec click using alternating polarity. About 3,000 sweeps were collected at a repetition rate of 20/sec The EEG was then epoched (0–40 msec), bandpass filtered (90–2000 Hz), and averaged in the time domain to derive ABRs for each participant.

Response Analysis and Quantification

ABR: ABR peak amplitudes and corresponding latencies were measured in response time waveforms as the peak positivity in the 5- to 9-msec time window.

FFR: FFR amplitudes were measured from the steady-state portion of the response via its fast Fourier transform (Song et al, 2011; Hornickel et al, 2012; Bidelman, 2015a). Spectral magnitude was measured as the Fourier bin corresponding to the fundamental frequency (F0) of the stimulus (here, 100 Hz). F0 amplitude provided an overall measure of the strength of the sustained following response. FFR onset latency was estimated from each FFR by first cross-correlating each response time waveform with the corresponding evoking stimulus (Galbraith and Brown, 1990; Bidelman, Villafuerte, et al, 2014). This provided a running correlation as a function of the lag between stimulus and response traces. The lag within a search window between 9 and 15 msec producing the maximum stimulus-to-response cross-correlation was taken as the onset latency for the brainstem FFR.

ERP: Peak amplitudes and latencies were measured for the prominent waves of the cortical ERPs (P1, N1, P2) within specific time intervals. Analysis windows were guided by visual inspection of the grand average ERP. P1 latency was taken as the peak positive deflection

between 45 and 70 msec; N1, the negativity between 100 and 110 msec; P2, the positivity between 145 and 155 msec.

Quantifying Inter- and Intrasubject Response Variability

Inter- (between) and intra- (within) subject variability was evaluated via ICC and coefficient of variation (CV), respectively. We computed the ICC to more directly compare the various classes of AEPs in terms of their consistency (i.e., repeatability). The ICC is a normalized statistic akin to a correlation coefficient that extends to multiple (i.e., >2) observations of the same unit or group (Koch, 1982; McGraw and Wong, 1996). The ICC allowed us to quantify the degree to which each AEP measure was consistent across the four test sessions and is a common metric to assess test–retest agreement between AEP waveforms (e.g., Tremblay et al, 2003). The ICC is advantageous here because it allowed us to directly compare response intersubject repeatability (a) using a singular statistic and (b) a metric that accounts for differences in the absolute scale between AEP classes (e.g., millisecond [ABR/FFR] versus decamillisecond [ERP] scale).

Because the ICC is only applicable to measuring intersubject variability, this metric could not directly assess response variability across test sessions at the individual subject level. To directly assess intrasubject variation "within each subject," we computed the CV across their individual test sessions (Dalebout and Robey, 1997). CV was calculated as CV = 100 × σ/μ, where σ and μ are the standard deviation and mean of the listener's response over the four test sessions. CVs were computed for each listener per response metric. Akin to ICC, CV is a normalized measure of dispersion which removes differences in scale between response indices. CVs allowed us to directly compare the intra- (within-) subject variability both between AEP metrics and levels of the pathway (i.e., brainstem versus cortex).

RESULTS

Brainstem (ABR, FFR) and cortical ERP waveforms are shown for the grand average and a representative subject in Figures 1A and 1B, respectively. Amplitude and latency metrics measured across test sessions are shown in Figure 2. Generally speaking, evoked potentials showed highly consistent patterns within and between subjects across repeated recordings regardless of AEP type. We first evaluated whether evoked response measures differed across test sessions via mixed model ANOVAs, with session as a single fixed factor (four levels) and subjects modeled as a random effect. We found no appreciable change in any of the subcortical response measures [ABR latency: $F_{(3,21)} = 0.15$, $p = 0.93$; ABR amplitude: $F_{(3,21)} = 1.27$, $p = 0.31$; FFR latency: $F_{(3,21)} = 0.69$, $p = 0.56$; FFR amplitude: $F_{(3,21)} = 1.12$, $p = 0.36$]. These results indicate both click-evoked and speech-evoked brainstem responses showed no appreciable change across test sessions.

In contrast to subcortical measures, the P1 wave of the cortical ERPs was modulated across test sessions becoming slightly prolonged (~6 msec) from the first to fourth session. This was confirmed by a linear effect of session on P1 latency [$F_{(3,21)} = 2.57$, $p = 0.0178$]. However, no

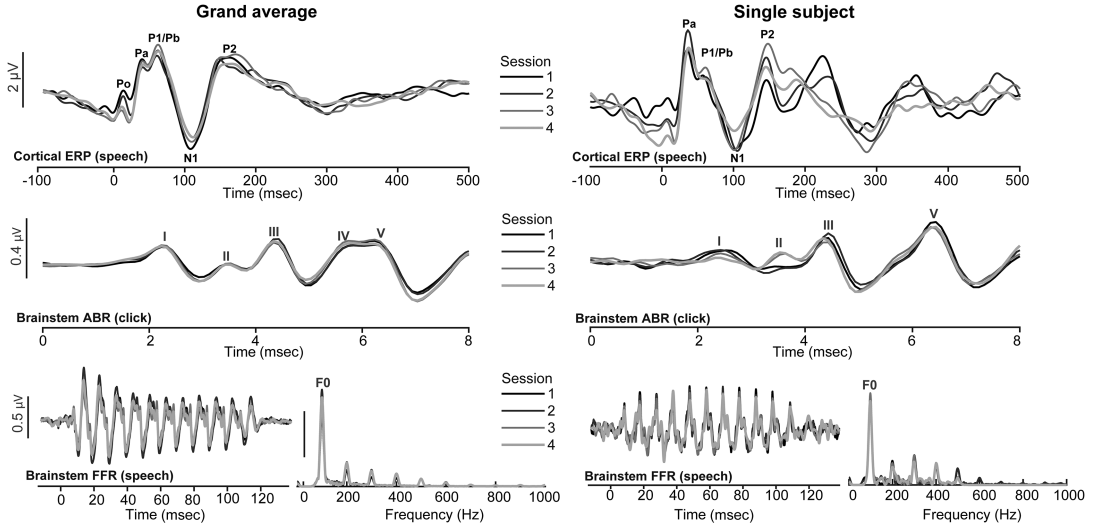

Figure 1. Brainstem and cortical AEP waveforms across test–retest sessions. (A) Grand average responses and (B) traces from a representative participant. (Bottom) Brainstem FFR time waveforms (left panel) and spectra (right panel) to speech, (middle) click-ABR, and (top) cortical ERPs to speech. Note the difference in the time scale (abscissa) and scalebar (ordinate) between response classes. Little variation is seen in the subcortical potentials (FFR, ABR) across the four test sessions. In contrast, cortical ERPs show more inherent inter- and intrasubject variability.

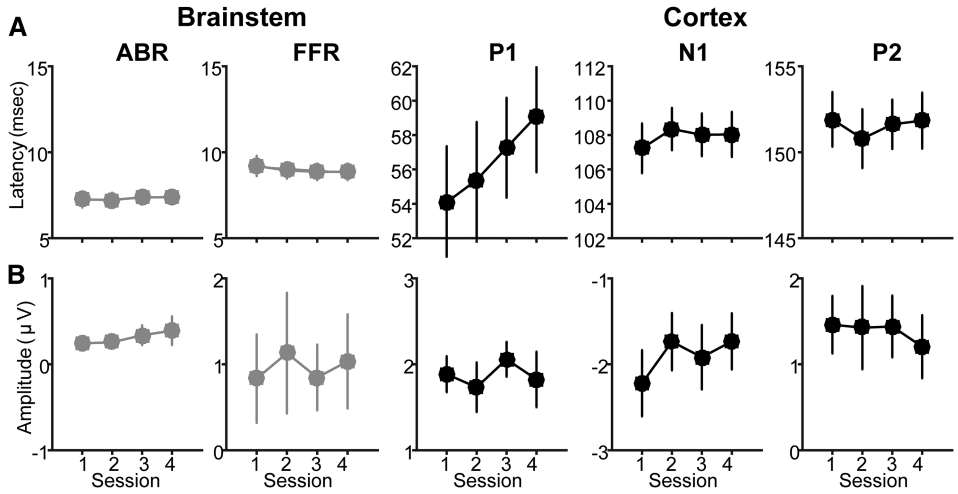

Figure 2. Intersession changes in brainstem and cortical response (A) latencies and (B) amplitudes. To facilitate comparisons, all latency and amplitude scales are identical across measures: 10 msec and 2 μV, respectively. Note that FFR F0 amplitude is measured in the spectral rather than the time domain as for other amplitude measures. Only cortical P1 latencies showed significant variation across sessions. Error bars = ±SEM.

other cortical wave latency nor amplitude characteristic changed across sessions [P1 amplitude: $F_{(3,21)} = 0.68$, $p = 0.57$; N1 latency: $F_{(3,21)} = 1.86$, $p = 0.16$; N1 amplitude: $F_{(3,21)} = 1.33$, $p = 0.29$; P2 latency: $F_{(3,21)} = 0.67$ $p = 0.57$; P2 amplitude: $F_{(3,21)} = 0.40$, $p = 0.76$]. Collectively, these findings suggest that the ABR, FFR, and cortical N1 and P2 measures were highly stable across multiple testing sessions across the period of a month. Only the P1 showed any systematic change across sessions. We return to this point in the "Discussion" section.

ICCs, indexing "intersubject" variability, are shown for each of the AEP measures in Figure 3A. All ICCs were significant at the $p < 0.0001$ level. Descriptive labels demarcate poor, fair, moderate, and strong consistency adopted from normal conventions for the ICC (Koch, 1982; McGraw and Wong, 1996). For brainstem measures, we found that click-ABR latency achieved strong (ICC = 0.76) test–retest reliability. In contrast, ABR amplitude showed more moderate repeatability (ICC = 0.65) across test sessions. Similarly, FFR latency (ICC = 0.86) and FFR amplitude (ICC = 0.94) both showed strong repeatability. It is of interest to note that FFR measures, on average, yielded higher intersubject test–retest than their conventional click-ABR counterparts.

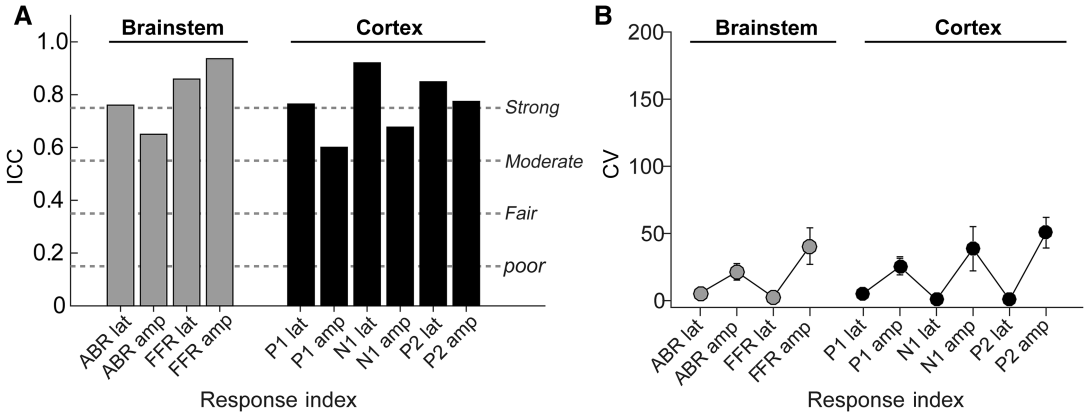

Figure 3. Inter- and intrasubject variability in brainstem and cortical AEP measures. (A) ICCs for brainstem and cortical response measured across test–retest sessions quantify intersubject variability between listeners. Descriptive labels (poor, fair, moderate, and strong) are adopted from normal conventions for the ICC (Koch, 1982; McGraw and Wong, 1996). Brainstem response measures show superior intersubject repeatability compared to most cortical measures. Declining ICC is observed for responses more central along the auditory neuroaxis (i.e., $ICC_{brainstem} > ICC_{cortical}$). Of the AEP response measures evaluated, FFR measures showed the highest test–retest consistency between subjects. (B) Intrasubject variability quantified via the CV of each listeners' responses across sessions. Although latency measures show less within-subject dispersion than amplitude measures, there is no systematic difference in within-subject variability between brainstem and cortical levels. Error bars = ±SEM. amp = amplitude; lat = latency.

In contrast to brainstem measures, cortical ERP metrics showed weaker ICCs across the board. However, lower ICCs for cortical responses might be expected given their higher within- and between-subject variability compared to brainstem potentials (Thornton and Coleman, 1975; Picton et al, 1978; Lauter and Karzon, 1990a,b; Bidelman, 2015a). Nevertheless, cortical P1-N1-P2 amplitude and latencies did show moderate (ICCs >0.55) to strong (ICCs >0.75) reliability. Broadly speaking, cortical ERP "latency" measures showed superior reliability across test sessions compared to amplitude. Of the obligatory waves, N1 latency showed the highest ICC of 0.93.

Mean "intra- (within-)subject" variability across subjects, as expressed by their CV across test sessions, is shown in Figure 3B. In agreement with previous work on nonspeech AEPs (Lauter and Karzon, 1990a,b; Dalebout and Robey, 1997), responses were remarkably stable within individual listeners, consistent with the notion that subjects' responses look "more like themselves than like each other" (Dalebout and Robey, 1997). Despite the small dispersion in response variance, intrasubject variability nevertheless differed across the AEP metrics [$F_{(9,63)} = 5.35, p < 0.0001$]. Bonferroni-adjusted contrasts revealed this effect was due to latency metrics being more stable (i.e., smaller CV) within subjects than their amplitude counterparts [$t_{(63)} = -6.27, p < 0.0001$]. However, intrasubject variability did not vary between brainstem and cortical levels for either latency [$t_{(63)} = -0.47, p = 1.00$] or amplitude [$t_{(63)} = 2.18, p = 0.13$] measures. This suggests that opposite to intersubject variability, where cortical AEPs were more variable than brainstem AEPs between listeners (i.e., Figure 3A), intrasubject variability (within single ears) does not become progressively greater at higher levels of the auditory system (cf. Dalebout and Robey, 1997).

Collectively, these findings suggest that brainstem AEPs (ABRs, FFRs) are inherently more consistent/reliable between subjects than their cortical counterparts, but show similar variation at the single subject level. Moreover, we find that the FFR appears to be the most reliable brain response (in terms of intersubject variability) across the subcortical and cortical AEP classes that were considered.

DISCUSSION

The present study assessed the stability of brainstem and cortical AEPs across multiple (×4) test sessions during a 1-mo period. We aimed to characterize and directly compare the test–retest reliability of subcortical and cortical classes of the AEPs "within the same set of listeners." Results show that both brainstem (ABR, FFR) and cortical (ERP: P1-N1-P2) auditory neural responses are highly repeatable between listeners, ranging from moderate to strong consistency (i.e., intraclass correlation). In particular, of the response measures assessed, ABR latency, FFR latency/amplitude, P1-N1-P2 latencies, and P2 amplitude showed the highest intersubject repeatability (ICC ≥0.75). Intrasubject assessment (i.e., CV) showed that amplitude measures were more variable than latency measures within single ears, but that responses had similar relative variability between brainstem and cortex at the individual subject level.

Repeatability of Speech-Evoked Auditory Cortical ERPs

The current study corroborates several previous attempts to evaluate repeatability of the various cortical AEP components over periods ranging from several weeks to years (McEvoy et al, 2000; Walhovd and Fjell, 2002; Tremblay et al, 2003; Tervaniemi et al, 2005; Williams et al, 2005; Hall et al, 2006; Cassidy et al, 2012; Huffmeijer et al, 2014; McFadden et al, 2014). Our data here confirm and extend these previous studies by similarly revealing moderate-to-strong reliability in the early cortical evoked potentials across multiple (×4) recording sessions over the course of a month both between and within subjects. Notably, we found that variability within single ears was exceptionally low. Overall speech-evoked ERPs were highly stable, showing intrasubject variability (i.e., CV ≈ 40–50) at or below that reported for ERPs elicited by nonspeech stimuli (Dalebout and Robey, 1997).

Evoked potential amplitude is known to vary with nonbiological factors (e.g., electrode impedance, orientation relative to source generators). This implies "prima facie" that amplitude might be a poor metric to reliably assess changes in the ERPs with certain experimental manipulations (e.g., training/learning, aging). Nevertheless, previous comparisons between measures suggest that early ERP amplitudes (P1 and N1) offer a stable indicator of cortical activity and can sometimes even achieve higher repeatability than latency metrics (Walhovd and Fjell, 2002; Tervaniemi et al, 2005; Cassidy et al, 2012). Our data here contrast previous results and suggest better repeatability of latency compared to amplitude measures of the cortical ERPs both between (inter-) and within (intra-) subject levels. Indeed, in all cases, P1-N1-P2 latency ICCs exceeded 0.75, indicating strong repeatability. In contrast, corresponding amplitude measures achieved ICCs that indicated only moderate repeatability. Similar findings were observed at the single subject level, where latency measures showed lower variance dispersion than amplitudes.

In this regard, it is noteworthy that of the cortical (and brainstem) ERP variables we evaluated, only P1 latency showed significant modulation across sessions. The magnitude of this latency change was small (6 msec) but is

nevertheless reliable in light of its high internal response consistency (i.e., ICC = 0.76). The neural mechanisms of this P1 effect are unknown. However, given that our cohort consisted entirely of females tested over the course of 1 mo, it is conceivable that P1 latency shifts could be related to the effects of hormonal changes during the menstrual cycle. Indeed, previous studies have reported peak latency shifts in auditory brainstem (Elkind-Hirsch et al, 1992) and cortical (Yadav et al, 2003) potentials over the menstrual cycle. Prolonged latencies may be related to estradiol-induced changes in inhibition resulting in decreased neural conduction velocity within the central auditory pathways and a delay of the AEPs. It is unclear why such latency effects were isolated to the P1 and not observed, for example, in early (ABR) or later (N1 and P2) waves of the auditory-evoked field. However, it has been suggested that central pathways are more influenced by hormone levels than peripheral structures (Elkind-Hirsch et al, 1992; Yadav et al, 2003). Moreover, estrogen and progesterone exert their strongest influence on neurotransmitter regulation in the diencephalon (e.g., thalamus) (McEwen et al, 1979), structures thought to contribute to the functional generation of the P1 wave (Scherg et al, 1989; Picton et al, 1999). This may account for why we observed latency shifts circumscribed to the cortical P1 rather than earlier (ABR/FFR) or later (N1 and P2) response components. Under this interpretation, we would probably not expect a P1 latency effect if our participants had included males. P1 is typically poorly defined and often difficult to measure at the individual subject level (Dalebout and Robey, 1997; Alain et al, 2013; Bidelman, Villafuerte et al, 2014; Bidelman and Alain, 2015). Consequently, we suspect P1 would have shown little modulation across sessions in a different cohort of listeners.

Repeatability of ABR and FFR

We found that the intra- and intersubject variability in ABR latency was negligibly small (submillisecond), in agreement with previous studies (Edwards et al, 1982; Oyler et al, 1991) and the precision of this measure for clinical hearing testing (Hall, 1992; Picton, 2010). Similarly, brainstem FFR latency and amplitudes were remarkably stable and showed no appreciable change across sessions. Hoormann et al (1992) assessed normal variation in response properties of tone-evoked FFRs (e.g., harmonic distribution and frequency-dependent amplitudes) but not did assess repeatability, per se. Moreover, few studies have assessed reliability of the "speech-evoked" FFR (Song et al, 2011; Hornickel et al, 2012). In their assessment of a variety of FFR response metrics, Hornickel et al (2012) reported highly stable responses to clean and noise-degraded speech tested at two time points over the period of 1 yr. While the magnitude of test–retest correlations varied considerably between response metrics ($r_{Spearman}$ = 0.12–0.82), FFR F0 amplitude showed test–retest correlations of $r_s \approx 0.80$, repeatability considered acceptable for clinical testing (Cicchetti, 1994). On the other hand, FFR latency measures were less stable, producing test–retest correlations of $r_s \leq 0.56$ (Hornickel et al, 2012) and raising concerns regarding the clinical utility of the FFR (e.g., McFarland and Cacace, 2012).

Our results corroborate and extend these previous findings by demonstrating robust repeatability of the speech-FFR in both amplitude and timing characteristics. Using multiple (×4) testing sessions allowed us to further confirm FFR repeatability and extend prior findings to more than a single pair of test sessions (e.g., Song et al, 2011; Hornickel et al, 2012). Both FFR latency and amplitude measures showed strong repeatability with ICCs of 0.86 and 0.94, respectively. Indeed, FFR amplitude measures demonstrated the strongest test–retest reliability among all 10 variables that were assessed across brainstem and cortical AEP classes.

Previous work has suggested that the brainstem FFR is distinct in its response characteristics from the more conventional click-ABR, differing in rate susceptibility (Krizman et al, 2010), frequency specificity (Picton et al, 1976, p. 105), spectral content (Bidelman, 2015b), susceptibility to noise masking (Cunningham et al, 2002; Russo et al, 2004), and latency-intensity changes (Akhoun et al, 2008). The functional distinction between ABR and FFR is further supported by recent brainstem studies in children, which reveal that the FFR provides superior diagnostic utility to the ABR in identifying auditory processing disorders and specific language impairments in this population (e.g., Rocha-Muniz et al, 2014). In light of its high stability within and between ears we observe here and functional distinctions reported elsewhere, we infer that the speech-FFR might provide a useful tool to augment current clinical assessment (i.e., ABR) in the objective evaluation of central auditory function (cf. Hornickel et al, 2012).

CONCLUSIONS

We recorded brainstem (ABR/FFR) and cortical AEPs (ERPs) in four test sessions over the course of a month to compare repeatability of subcortical versus cortical classes of brain activity within the same set of listeners. With few exceptions (P1 latency), our findings show that auditory brainstem and cortical amplitude/latency measures are remarkably stable across listeners; no significant variation was observed in response measures. ICCs further revealed that speech-evoked FFR amplitude and latency measures achieved superior repeatability (ICC >0.8) among the more widely used brainstem (ABR) and cortical (P1-N1-P2) AEPs. Intrasubject measures (CV) revealed that while latencies were more stable than amplitudes, brainstem

and cortical AEPs did not differ in their variability at the single subject level. Our data suggest that the hierarchy of AEPs show similar relative variability "within" individuals (brainstem ≈ cortex) but become progressively more variable "between" subjects along the ascending neuroaxis (cortex > brainstem), possibly highlighting differences in individual listeners' experience(s) and/or suprathreshold auditory skills (cf. Chandrasekaran and Kraus, 2010; Bidelman, 2017; Kraus et al, 2017).

REFERENCES

Akhoun I, Gallégo S, Moulin A, Ménard M, Veuillet E, Berger-Vachon C, Collet L, Thai-Van H. (2008) The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme /ba/ in normal-hearing adults. *Clin Neurophysiol* 119(4):922–933.

Alain C, Roye A, Arnott SR. (2013) Middle- and long-latency auditory evoked potentials: What are they telling us on central auditory disorders? In: Celesia GG, ed. *Handbook of Clinical Neurophysiology: Disorders of Peripheral and Central Auditory Processing*. The Netherlands: Elsevier.

Banai K, Abrams D, Kraus N. (2007) Sensory-based learning disability: insights from brainstem processing of speech sounds. *Int J Audiol* 46(9):524–532.

Banai K, Hornickel J, Skoe E, Nicol T, Zecker S, Kraus N. (2009) Reading and subcortical auditory function. *Cereb Cortex* 19(11): 2699–2707.

Basu M, Krishnan A, Weber-Fox C. (2010) Brainstem correlates of temporal auditory processing in children with specific language impairment. *Dev Sci* 13(1):77–91.

Bidelman GM. (2014) Objective information-theoretic algorithm for detecting brainstem-evoked responses to complex stimuli. *J Am Acad Audiol* 25(8):715–726.

Bidelman GM, Moreno S, Alain C. (2013) Tracing the emergence of categorical speech perception in the human auditory system. *Neuroimage* 79:201–212.

Bidelman GM. (2015a) Towards an optimal paradigm for simultaneously recording cortical and brainstem auditory evoked potentials. *J Neurosci Methods* 241:94–100.

Bidelman GM. (2015b) Multichannel recordings of the human brainstem frequency-following response: scalp topography, source generators, and distinctions from the transient ABR. *Hear Res* 323:68–80.

Bidelman GM. (2017) Communicating in challenging environments: noise and reverberation. In: Kraus N, Anderson S, White-Schwoch T, Fay RR, Popper AN, eds. *Springer Handbook of Auditory Research: The Frequency-Following Response: A Window into Human Communication*. New York, NY: Springer Nature.

Bidelman GM, Alain C. (2015) Musical training orchestrates coordinated neuroplasticity in auditory brainstem and cortex to counteract age-related declines in categorical vowel perception. *J Neurosci* 35(3):1240–1249.

Bidelman GM, Villafuerte JW, Moreno S, Alain C. (2014) Age-related changes in the subcortical-cortical encoding and categorical perception of speech. *Neurobiol Aging* 35(11):2526–2540.

Bidelman GM, Weiss MW, Moreno S, Alain C. (2014) Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. *Eur J Neurosci* 40(4):2662–2673.

Bornstein MH, Rothstein H, Cohen J. (1997) Power and Precision™ (v1) [Computer software], Biostat, Englewood, NJ.

Caruso S, Maiolino L, Rugolo S, Intelisano G, Farina M, Cocuzza S, Serra A. (2003) Auditory brainstem response in premenopausal women taking oral contraceptives. *Hum Reprod* 18:(1):85–89.

Cassidy SM, Robertson IH, O'Connell RG. (2012) Retest reliability of event-related potentials: evidence from a variety of paradigms. *Psychophysiology* 49(5):659–664.

Chandrasekaran B, Kraus N. (2010) The scalp-recorded brainstem response to speech: neural origins and plasticity. *Psychophysiology* 47(2):236–246.

Cicchetti DV. (1994) Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol Assess* 6:284–290.

Cunningham J, Nicol T, King C, Zecker SG, Kraus N. (2002) Effects of noise and cue enhancement on neural responses to speech in auditory midbrain, thalamus and cortex. *Hear Res* 169(1–2): 97–111.

Dalebout SD, Robey RR. (1997) Comparison of the intersubject and intrasubject variability of exogenous and endogenous auditory evoked potentials. *J Am Acad Audiol* 8(5): 342–354.

Edwards RM, Buchwald JS, Tanguay PE, Schwafel JA. (1982) Sources of variability in auditory brain stem evoked potential measures over time. *Electroencephalogr Clin Neurophysiol* 53 (2):125–132.

Elkind-Hirsch KE, Stoner WR, Stach BA, Jerger JF. (1992) Estrogen influences auditory brainstem responses during the normal menstrual cycle. *Hear Res* 60(2):143–148.

Galbraith GC, Arbagey PW, Branski R, Comerci N, Rector PM. (1995) Intelligible speech encoded in the human brain stem frequency-following response. *Neuroreport* 6(17):2363–2367.

Galbraith GC, Brown WS. (1990) Cross-correlation and latency compensation analysis of click-evoked and frequency-following brain-stem responses in man. *Electroencephalogr Clin Neurophysiol* 77(4):295–308.

Hall JW. (1992) *Handbook of Auditory Evoked Responses*. Needham Heights, MA: Allyn and Bacon.

Hall MH, Schulze K, Rijsdijk F, Picchioni M, Ettinger U, Bramon E, Freedman R, Murray RM, Sham P. (2006) Heritability and reliability of P300, P50 and duration mismatch negativity. *Behav Genet* 36(6):845–857.

Hillyard SA, Picton TW. (1979) Event-related brain potentials and selective information processing in man. In: Desmedt JE, ed. *Progress in Clinical Neurophysiology*. Basel, Switzerland: Karger.

Hoormann J, Falkenstein M, Hohnsbein J, Blanke L. (1992) The human frequency-following response (FFR): normal variability and relation to the click-evoked brainstem response. *Hear Res* 59(2):179–188.

Hornickel J, Knowles E, Kraus N. (2012) Test-retest consistency of speech-evoked auditory brainstem responses in typically-developing children. *Hear Res* 284(1–2):52–58.

Hu L, Mouraux A, Hu Y, Iannetti GD. (2010) A novel approach for enhancing the signal-to-noise ratio and detecting automatically event-related potentials (ERPs) in single trials. *Neuroimage* 50:(1):99–111.

Huffmeijer R, Bakermans-Kranenburg MJ, Alink LRA, van Ijzendoorn MH. (2014) Reliability of event-related potentials: the influence of number of trials and electrodes. *Physiol Behav* 130:13–22.

Jerger J, Jerger S. (1985) Audiologic applications of early, middle, and late auditory evoked potentials. *Hear J* 38:31–36.

Jerger J, Musiek F. (2000) Report of the consensus conference on the diagnosis of auditory processing disorders in school-aged children. *J Am Acad Audiol* 11(9):467–474.

Klatt DH, Klatt LC. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am* 87(2):820–857.

Koch GG. (1982) Intraclass correlation coefficient. In: Kotz S, Johnson NL, eds. *Encyclopedia of Statistical Sciences*. New York, NY: John Wiley & Sons.

Kraus N, Anderson S, White-Schwoch T, Fay RR, Popper AN. (2017) *The Frequency-Following Response: A Window into Human Communication*. New York, NY: Springer Nature.

Krishnan A. (2007) Human frequency following response. In: Burkard RF, Don M, Eggermont JJ, eds. *Auditory Evoked Potentials: Basic Principles and Clinical Application*. Baltimore, MD: Lippincott Williams & Wilkins.

Krizman JL, Skoe E, Kraus N. (2010) Stimulus rate and subcortical auditory processing of speech. *Audiol Neurootol* 15(5): 332–342.

Lauter JL, Karzon RG. (1990a) Individual differences in auditory electric responses: comparisons of between-subject and within-subject variability. V. Amplitude-variability comparisons in early, middle, and late responses. *Scand Audiol* 19(4):201–206.

Lauter JL, Karzon RG. (1990b) Individual differences in auditory electric responses: comparisons of between-subject and within-subject variability. IV. Latency-variability comparisons in early, middle, and late responses. *Scand Audiol* 19(3):175–182.

Lightfoot G, Kennedy V. (2006) Cortical electric response audiometry hearing threshold estimation: accuracy, speed, and the effects of stimulus presentation features. *Ear Hear* 27(5):443–456.

McEvoy LK, Smith ME, Gevins A. (2000) Test-retest reliability of cognitive EEG. *Clin Neurophysiol* 111(3):457–463.

McEwen BS, Davis PG, Parsons B, Pfaff DW. (1979) The brain as a target for steroid hormone action. *Annu Rev Neurosci* 2:65–112.

McFadden KL, Steinmetz SE, Carroll AM, Simon ST, Wallace A, Rojas DC. (2014) Test-retest reliability of the 40 Hz EEG auditory steady-state response. *PLoS One* 9(1):e85748.

McFarland DJ, Cacace AT. (2012) Questionable reliability of the speech-evoked auditory brainstem response (sABR) in typically-developing children. *Hear Res* 287(1–2):1–2, author reply 3–5.

McGraw KO, Wong SP. (1996) Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1:30–46.

Musacchia G, Strait D, Kraus N. (2008) Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hear Res* 241(1–2):34–42.

Näätänen R. (1992) *Attention and Brain Function*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Okamoto H, Stracke H, Bermudez P, Pantev C. (2011) Sound processing hierarchy within human auditory cortex. *J Cogn Neurosci* 23(8):1855–1863.

Oyler RF, Lauter JL, Matkin ND. (1991) Intrasubject variability in the absolute latency of the auditory brainstem response. *J Am Acad Audiol* 2(4):206–213.

Picton TW. (2010) *Human Auditory Evoked Potentials*. 1st ed. San Diego, CA: Plural Publishing.

Picton TW, Alain C, Woods DL, John MS, Scherg M, Valdes-Sosa P, Bosch-Bayard J, Trujillo NJ. (1999) Intracerebral sources of human auditory-evoked potentials. *Audiol Neurootol* 4(2):64–79.

Picton TW, Hillyard SA. (1974) Human auditory evoked potentials. II. Effects of attention. *Electroencephalogr Clin Neurophysiol* 36(2):191–200.

Picton TW, Hillyard SA, Galambos R, Schiff M. (1971) Human auditory attention: a central or peripheral process? *Science* 173:(3994):351–353.

Picton TW, Woods DL, Baribeau-Braun J, Healey TM. (1976) Evoked potential audiometry. *J Otolaryngol* 6(2):90–119.

Picton TW, Woods DL, Proulx GB. (1978) Human auditory sustained potentials. II. Stimulus relationships. *Electroencephalogr Clin Neurophysiol* 45(2):198–210.

Ponton CW, Eggermont JJ, Kwong B, Don M. (2000) Maturation of human central auditory system activity: evidence from multi-channel evoked potentials. *Clin Neurophysiol* 111(2):220–236.

Rocha-Muniz CN, Befi-Lopes DM, Schochat E. (2012) Investigation of auditory processing disorder and language impairment using the speech-evoked auditory brainstem response. *Hear Res* 294:(1–2):143–152.

Rocha-Muniz CN, Befi-Lopes DM, Schochat E. (2014) Sensitivity, specificity and efficiency of speech-evoked ABR. *Hear Res* 317:15–22.

Russo N, Nicol T, Musacchia G, Kraus N. (2004) Brainstem responses to speech syllables. *Clin Neurophysiol* 115(9):2021–2030.

Salinsky MC, Oken BS, Morehead L. (1991) Test-retest reliability in EEG frequency analysis. *Electroencephalogr Clin Neurophysiol* 79(5):382–392.

Scherg M, Vajsar J, Picton TW. (1989) A source analysis of the late human auditory evoked potentials. *J Cogn Neurosci* 1(4):336–355.

Sininger YS, Abdala C, Cone-Wesson B. (1997) Auditory threshold sensitivity of the human neonate as measured by the auditory brainstem response. *Hear Res* 104(1–2):27–38.

Skoe E, Kraus N. (2010) Auditory brain stem response to complex sounds: a tutorial. *Ear Hear* 31(3):302–324.

Song JH, Banai K, Russo NM, Kraus N. (2006) On the relationship between speech- and nonspeech-evoked auditory brainstem responses. *Audiol Neurootol* 11(4):233–241.

Song JH, Nicol T, Kraus N. (2011) Test-retest reliability of the speech-evoked auditory brainstem response. *Clin Neurophysiol* 122(2):346–355.

Stapells DR. (2000) Threshold estimation by the tone-evoked auditory brainstem response: a literature meta-analysis. *J Speech Lang Pathol Audiol* 42:74–83.

Starr A, Picton TW, Sininger Y, Hood LJ, Berlin CI. (1996) Auditory neuropathy. *Brain* 119(Pt 3):741–753.

Tervaniemi M, Sinkkonen J, Virtanen J, Kallio J, Ilmoniemi RJ, Salonen O, Näätänen R. (2005) Test-retest stability of the magnetic mismatch response (MMNm). *Clin Neurophysiol* 116(8):1897–1905.

Thornton ARD, Coleman MJ. (1975) The adaptation of cochlear and brainstem auditory evoked potentials in humans. *Electroencephalogr Clin Neurophysiol* 39(4):399–406.

Tremblay KL, Friesen L, Martin BA, Wright R. (2003) Test-retest reliability of cortical evoked potentials using naturally produced speech sounds. *Ear Hear* 24(3):225–232.

Walhovd KB, Fjell AM. (2002) One-year test-retest reliability of auditory ERPs in young and old adults. *Int J Psychophysiol* 46:(1):29–40.

Weiss MW, Bidelman GM. (2015) Listening to the brainstem: musicianship enhances intelligibility of subcortical representations for speech. *J Neurosci* 35(4):1687–1691.

Williams LM, Simms E, Clark CR, Paul RH, Rowe D, Gordon E. (2005) The test-retest reliability of a standardized neurocognitive and neurophysiological test battery: "neuromarker." *Int J Neurosci*.115(12):1605-1630.

Woods DL, Hillyard SA. (1978) Attention at the cocktail party: brainstem evoked responses reveal no peripheral gating. In: Otto DA, ed. *Multidisciplinary Perspectives in Event-Related Brain Potential Research (EPA 600/9-77-043)*. Washington, DC: U.S. Government Printing Office.

Yadav A, Tandon OP, Vaney N. (2003) Long latency auditory evoked responses in ovulatory and anovulatory menstrual cycle. *Indian J Physiol Pharmacol* 47(2):179–184.