Research Article

Reliability of Measures of N1 Peak Amplitude of the Compound Action Potential in Younger and Older Adults

Carolyn M. McClaskey,^a James W. Dias,^a Judy R. Dubno,^a and Kelly C. Harris^a

Purpose: Human auditory nerve (AN) activity estimated from the amplitude of the first prominent negative peak (N1) of the compound action potential (CAP) is typically quantified using either a peak-to-peak measurement or a baseline-corrected measurement. However, the reliability of these 2 common measurement techniques has not been evaluated but is often assumed to be relatively poor, especially for older adults. To address this question, the current study (a) compared test–retest reliability of these 2 methods and (b) tested the extent to which measurement type affected the relationship between N1 amplitude and experimental factors related to the stimulus (higher and lower intensity levels) and participants (younger and older adults).

Method: Click-evoked CAPs were recorded in 24 younger (aged 18–30 years) and 20 older (aged 55–85 years) adults with clinically normal audiograms up to 3000 Hz. N1 peak amplitudes were estimated from peak-to-peak measurements (from N1 to P1) and baseline-corrected measurements for 2 stimulus levels (80 and 110 dB pSPL). Baseline-corrected measurements were made with 4 baseline windows. Each stimulus level was presented twice, and test–retest reliability of these 2 measures was assessed using the intraclass correlation coefficient. Linear mixed models were used to evaluate the extent to which age group and

click level uniquely predicted N1 amplitude and whether the predictive relationships differed between N1 measurement techniques.

Results: Both peak-to-peak and baseline-corrected measurements of N1 amplitude were found to have good-to-excellent reliability, with intraclass correlation coefficient values > 0.60. As expected, N1 amplitudes were significantly larger for younger participants compared with older participants for both measurement types and were significantly larger in response to clicks presented at 110 dB pSPL than at 80 dB pSPL for both measurement types. Furthermore, the choice of baseline window had no significant effect on N1 amplitudes using the baseline-corrected method.

Conclusions: Our results suggest that measurements of AN activity can be robustly and reliably recorded in both younger and older adults using either peak-to-peak or baseline-corrected measurements of the N1 of the CAP. Peak-to-peak measurements yield larger N1 response amplitudes and are the default measurement type for many clinical systems, whereas baseline-corrected measurements are computationally simpler. Furthermore, the relationships between AN activity and stimulus- and participant-related variables were not affected by measurement technique, which suggests that these relationships can be compared across studies using different techniques for measuring the CAP N1.

Electrocochleography (ECoChG) is a useful tool for noninvasively measuring the function of the peripheral auditory system. ECoChG measures the electrical potentials of the cochlea and auditory nerve (AN) that are generated in response to acoustic stimulation. These potentials are measured from electrodes placed in the ear canal (Bramhall, Konrad-Martin, McMillan, & Griest,

2017; Liberman, Epstein, Cleveland, Wang, & Maison, 2016), on the tympanum (Chertoff, Lichtenhan, & Willis, 2010; Harris, Vaden, McClaskey, Dias, & Dubno, 2018; Lichtenhan & Chertoff, 2008; Murnane, Prieve, & Relkin, 1998), on the promontory (Eggermont, 1976a; Santarelli, Starr, Michalewski, & Arslan, 2008), or at the round window (Bourien et al., 2014; Earl & Chertoff, 2010; Eggermont, 1976a; Kujawa & Liberman, 2009; Schmiedt, Mills, & Boettcher, 1996; Scott et al., 2016). Various aspects of the resulting ECoChG waveform, particularly the initial complex that reflects AN activity called the compound action potential (CAP), provide useful information for clinicians and researchers seeking to understand and diagnose peripheral auditory pathologies. Although some researchers have

^aMedical University of South Carolina, Charleston

Correspondence to Carolyn M. McClaskey: mcclaske@musc.edu

Editor-in-Chief: Frederick (Erick) Gallun

Editor: Steve Aiken

Received March 20, 2018

Revision received May 7, 2018

Accepted May 23, 2018

https://doi.org/10.1044/2018_JSLHR-H-18-0097

Disclosure: The authors have declared that no competing interests existed at the time of publication.

expressed concern that the CAP may not be a reliable assessment of AN activity, especially in older adults, and is too difficult to measure in humans (Mehraei, Gallardo, Shinn-Cunningham, & Dau, 2017; Mehraei et al., 2016; Plack et al., 2016), the CAP represents the most direct measure of AN activity and, as shown here, is robust and reliable in both younger and older adults.

The first prominent negative peak of the CAP (N1) represents the summed synchronous firing of AN fibers. N1 amplitude and latency, along with other characteristics (Harris et al., 2018), can serve as estimates of AN structure and function. Historically, the amplitude of the peak was quantified by calculating the voltage difference between the N1 and any one of several surrounding positive peaks (Bourien et al., 2014; Chertoff et al., 2010; Murnane et al., 1998; Schwartz-Leyzac & Pfingst, 2016), a technique referred to here as *peak-to-peak amplitude measurement* (see Figure 1). These reference peaks are most commonly the positive peak just after the N1 (P1), or the positive inflection point just before the N1, which is sometimes also called the cochlear summing potential (SP). However, because the SP is a mechanical potential generated primarily by the hair cells and the N1 is a neural response generated by the AN, the specific design of the stimulus and apparatus can result in significant changes to the direction of the SP and in the latency of the N1 (Eggermont, 1976b, 2017). As such, the SP may not serve as an appropriate reference peak for peak-to-peak amplitude measurements. Instead, it may be useful for characterizing cochlear synaptopathy, auditory neuropathy, and Ménière's disease (Baba et al., 2009; Eggermont, 1976b, 2017; Gibson, Moffat, & Ramsden, 1977; Liberman et al., 2016; Santarelli & Arslan, 2002).

Regardless of the choice of reference peak, peak-to-peak measurements require the identification of two peaks rather than one, and variability in the responses of the surrounding peaks can introduce additional variability to the measurement of N1 amplitude.

An alternative method is to measure N1 amplitude as an absolute voltage relative to a time-averaged baseline, referred to here as a *baseline-corrected amplitude measurement* (see Figure 1). In this method, the average amplitude of a portion of the CAP waveform that occurs before the eliciting stimulus is subtracted from the entire CAP waveform before the voltage of the peaks is measured. Such baseline corrections are used widely in cortical electroencephalography (EEG), where a longer time course allows for more stable baseline periods. Peak-to-peak measurements are also considered undesirable in EEG because subsequent peaks/components are often separated by tens of milliseconds and have different cortical generators (M. X. Cohen, 2014; Luck, 2014). Baseline-corrected amplitude measurements are computationally simpler than peak-to-peak amplitude measurements, requiring the identification of only a single peak in the waveform. For these reasons, baseline-corrected amplitude measurements have become increasingly common in ECoG studies (e.g., Harris et al., 2018; Parthasarathy, Datta, Torres, Hopkins, & Bartlett, 2014; Scott et al., 2016).

Previous research has shown that the choice of baseline correction and other preprocessing parameters, such as filtering and artifact rejection, can affect poststimulus response characteristics (M. X. Cohen, 2014; Hu, Xiao, Zhang, Mouraux, & Iannetti, 2014; Luck, 2014; Tanner, Norton, Morgan-Short, & Luck, 2016). More specifically, the choice of a baseline window can affect poststimulus

Figure 1. Compound action potential (CAP) waveform (Solid black line) in response to a click train presented at 110 dB pSPL averaged across all younger participants. Light gray halo indicates ± 1 SEM. Dark gray box indicates a 2-ms prestimulus baseline window. Dashed black horizontal line indicates the baseline, determined by averaging the amplitude of an individual waveform within the baseline window and subtracting it from the response. The N1 and P1 peaks are labeled. Baseline-corrected and peak-to-peak N1 amplitude measurement methods are indicated by vertical black lines.

response variability and amplitude. EEG researchers typically select a prestimulus baseline a few hundred milliseconds in duration or approximately 20% of the full epoch (M. X. Cohen, 2014; Luck, 2014). However, ideal baseline parameters may differ in ECoG, where epochs are 10–20 ms in duration rather than 500–2000 ms. The choice of baseline window is therefore an important consideration when using baseline-corrected amplitude measurements.

Although both peak-to-peak and baseline-corrected amplitude measurement techniques have been used to estimate N1 peak amplitude in the human CAP, the reliability of these measures is unknown but often assumed to be relatively poor, especially for older adults. The reliability of N1 amplitude measurements can have important implications for clinical diagnoses and conclusions drawn from CAP and AN research. Thus, the goal of the current study is to examine the test–retest reliability of peak-to-peak and peak-to-baseline N1 amplitude measurements in younger and older human listeners. We first assessed four baseline windows to determine if the choice of window affects test–retest reliability. We then assessed how age group and click level relate to N1 peak amplitude and examined the extent to which amplitude measurement type affects these relationships. We predicted that both measurement types would demonstrate robust test–retest reliability. Furthermore, we hypothesized that both age group and click level predict N1 amplitude independent of measurement type.

Method

Participants

Participants were 44 adults with clinically normal audiograms up to 3000 Hz recruited from the greater Charleston community: 24 younger adults (age: $M = 25$ years, $SD = 2.7$ years; 11 men, 13 women) and 20 older adults ($M = 65$ years, $SD = 6.5$ years; six men, 14 women). All participants had pure-tone air conduction thresholds at or better than 25 dB HL at 500, 1000, 2000, and 3000 Hz as well as between-ear threshold differences less than 15 dB (see Figure 2). All testing was done in the right ear. Participants had normal otoscopic assessments and no history of otologic or neurologic problems and provided informed written consent before testing. Testing was initiated after approval by the Medical University of South Carolina's Institutional Review Board.

Stimuli and Equipment

Experimental stimuli were click trains of 100- μ s rectangular pulses of alternating polarity presented at a rate of 11.1 clicks/s. Clicks were presented to the right ear through an ER-3C insert earphone (Etymotic Research, Inc., Elk Grove Village, IL) in blocks of approximately 1,100 presentations. Blocks were presented at stimulus intensities ranging from 70 to 110 dB pSPL in 10-dB steps. Click trains were presented twice at each level, and the order of blocks was counterbalanced and randomized, with the requirement that all five stimulus levels had to be presented

Figure 2. Mean pure-tone air conduction thresholds at audiometric frequencies for the right (test) ear in younger (filled triangles) and older (open circles) participants. Error bars indicate the range.

once before any level was repeated. Levels of 110 and 80 dB pSPL were used in the current analysis, whereas the remaining stimulus levels were used to assist in peak identification.

Data Collection

CAP responses were recorded using a tympanic membrane electrode (Sanibel) in the right ear, an inverting electrode placed on the contralateral (left) mastoid, and a low forehead grounding electrode. Auditory brainstem responses were simultaneously recorded for reference using a high forehead active electrode, an inverting mastoid electrode (on the right mastoid), and a low forehead grounding electrode (shared with CAP setup). All recordings were collected at a sampling rate of 20 kHz using a custom headstage (30x Gain, Tucker Davis Technologies, FL) connected to the bipolar channels of a Neuroscan SynAmpsRT amplifier (AC mode, 2010x Gain, Compumedics USA, NC). Testing was done in an acoustically and electrically shielded room. Participants reclined in a chair and were encouraged to sleep or rest quietly for the duration of testing.

Continuous data were bandpass filtered offline from 300 to 3000 Hz using an eighth-order Butterworth filter implemented in the EEGLab package (Delorme & Makeig, 2004) in MATLAB (The MathWorks, Inc., Natick, MA). Responses were epoched either from –2.6 ms prestimulus to +10.4 ms poststimulus (short epoch length) or from –11.6 ms prestimulus to +10.4 ms poststimulus (long epoch length) and then baseline corrected to a time-averaged window (see below) using the baseline subtraction method (M. X. Cohen, 2014).

Data were screened for artifacts using a moving window peak-to-peak threshold implemented in ERPLab (Lopez-Calderon & Luck, 2014). The window size was 2 ms, and the step size was 0.5 ms. An initial voltage threshold

was set to 45 μV and increased if necessary. For those participants in which 45 μV was inappropriate (unnecessarily rejected > 30% of the trials), a threshold was chosen to reject 2% of the trials and then increased if necessary. Visual inspection was used to ensure that only trials with excessive movement artifacts were rejected. An average of 7.1% of trials were rejected in each block. Artifact-free trials were averaged to yield the averaged CAP.

Baseline Window Comparisons

To allow for a comparison of different baseline corrections, four baseline windows were used. Both a long epoch (22 ms) and a short epoch (13 ms) were included, and both a prestimulus baseline and a whole-trial baseline window were used for each epoch length, denoted here as the following:

1. Long prestimulus baseline, from -11.6 to -0.6 ms [$-11.6, -0.6$]
2. Long whole-trial baseline, from -11.6 to $+10.4$ ms [$-11.6, +10.4$]
3. Short prestimulus baseline, from -2.6 to -0.6 ms [$-2.6, -0.6$]
4. Short whole-trial baseline, from -2.6 to $+10.4$ ms [$-2.6, +10.4$]

Both a prestimulus baseline and a whole-epoch baseline were included to provide a more thorough analysis of the effects of baseline window. A 2-ms window was chosen for the short prestimulus baseline because this is approximately 20% of the epoch duration (Luck, 2014), whereas a 10-ms window was chosen for the long prestimulus baseline because this is approximately equivalent to poststimulus epoch length. The prestimulus baseline period ended at -0.6 ms to prevent potential stimulus artifacts and poststimulus responses (Tanner et al., 2016) from contaminating the baseline control window.

N1 Amplitude Measurements

The N1 of the CAP was defined as the first prominent negative peak in the response that also fulfilled the following two constraints: The N1 (a) could not occur earlier than 1 ms poststimulus because of physiological constraints and (b) was repeatable in that it occurred in both presentations of each click level. Where necessary, the auditory brainstem response channel was used as a reference to help identify Waves I (corresponding to N1 of the CAP) and V. The P1 was defined as the positive peak after the N1.

Two independent and experienced reviewers visually confirmed all peak identifications, with one reviewer blind to the experimental groups and conditions. A third independent experienced reviewer was consulted if necessary. Conflicting peak identifications were jointly resolved before data analysis. The peak amplitude of the N1 was then measured as the peak-to-peak voltage difference from the N1 to the P1 (peak-to-peak amplitude) and as amplitude relative to baseline (baseline-corrected amplitude).

Data Analysis

Test-retest reliability was assessed using the intra-class correlation coefficient (ICC). ICCs are a special class of correlations that can be performed on unordered measurements, can be calculated between more than two measurements, and have the ability to evaluate both random and systematic errors (McGraw & Wong, 1996; Shrout & Fleiss, 1979; Weir, 2005). For this reason, ICCs are uniquely suited to assess test-retest reliability across trials and have several advantages over other types of correlations. For example, Pearson's r is limited to only two measurements, cannot detect systematic error, and is thus not recommended for test-retest reliability (Koo & Li, 2016; Weir, 2005), whereas Cronbach's alpha is designed to measure interrater reliability for large sample sizes, designates the lower limit of reliability, and is also not recommended (Sijtsma, 2009). There are several forms of the intraclass correlation, and the choice of which to use depends on the parameters of the model (one- vs. two-way, fixed vs. random effects), whether the final score is a single measure or an average of multiple scores (k), and whether relative reliability (consistency) or absolute reliability (absolute agreement) is being measured. Finally, two-way ICC models are ideal for evaluations of test-retest reliability (Koo & Li, 2016; Weir, 2005).

In the current study, test-retest reliability was assessed between the first and second presentations of each stimulus level using the two-way random ICC, ICC(2, k), which evaluates both random and systematic errors (McGraw & Wong, 1996; Shrout & Fleiss, 1979; Weir, 2005) and absolute reliability. Because standard ECoG methods average N1 amplitudes across multiple presentations, the averaged ICC criterion (k) is used. ICC values are spuriously low without adequate within-group variability; thus, ICC values were calculated across both younger and older participants for each condition. Pilot analyses showed that the standard error of measurement across both presentations was equivalent between younger and older participants (data not shown), supporting the use of ICC values calculated across all participants (Weir, 2005). ICC values less than 0.40 are considered to reflect poor reliability, values between 0.40 and 0.59 indicate fair reliability, values between 0.60 and 0.74 indicate good reliability, and values of 0.75 and higher indicate excellent reliability (Cicchetti, 1994; Hallgren, 2012).

To compare between different ICC values, ICCs were first transformed into z scores using Fisher's r -to- z transformation for correlations (Fisher, 1936). Z scores were calculated as follows:

$$z = \frac{1}{2} \ln \left(\frac{1 + \text{ICC}}{1 - \text{ICC}} \right) \quad (1)$$

The difference in z scores between two ICCs was then computed, and the associated probability was used to determine whether differences were significant at $\alpha = .05$ (Fisher, 1936; McGraw & Wong, 1996).

Separate linear mixed-effects models, using the lme4 package for R ($\times 64$ v3.1.2), were tested to determine the

extent to which age group and click level uniquely predicted peak-to-peak N1 amplitude and baseline-corrected N1 amplitude. Peak-to-peak and baseline-corrected measures are derived from the same CAP waveforms, and their N1 amplitude measurements are interdependent; thus, including them in a single linear mixed-effects model to test measurement type as a fixed factor violates the assumption of independence. As an alternate method of comparing the two measurement techniques and to determine the extent to which relationships between age group, click level, and N1 amplitude differed by measurement technique, the standardized coefficients for age and click level from the separate models were compared via *Z* tests. *Z* scores were calculated as follows:

$$z = \frac{\beta_1 - \beta_2}{\sqrt{SE_{\beta_1}^2 + SE_{\beta_2}^2}} \quad (2)$$

where β_1 and β_2 indicate the standardized regression coefficients for the two models being compared and SE_{β_1} and SE_{β_2} indicate the standard error for each model. The corresponding probabilities associated with the computed *z* scores were then used to determine whether differences were significant at $\alpha = .05$ (Clogg, Petkova, & Haritou, 1995; J. E. Cohen, Cohen, West, & Aiken, 2003; Paternoster, Brame, Mazerolle, & Piquero, 1998).

Results

Effects of Baseline Window on Baseline-Corrected Measurements

First, an analysis was conducted to determine the extent to which baseline window affects baseline-corrected N1 peak amplitudes and their test-retest reliability. We examined baseline-corrected measurements using four time-averaged baseline windows at 110 dB pSPL, as described earlier.

An N1 peak was present in both 110 dB pSPL blocks for all 44 participants. One older participant was excluded from the short prestimulus baseline condition [−2.6, −0.6] because the peak amplitude for one of the two presentations fell above 0 μ V. ICCs of the N1 amplitudes and 95% confidence intervals were calculated to determine the test-retest reliability of the four baseline window conditions. ICC(2,*k*) values are shown in Table 1. A comparison of ICC values via *z* tests showed no significant difference between the ICC values for the four baseline windows ($p > .05$). All ICC values indicated excellent test-retest reliability regardless of baseline window (Cicchetti, 1994).

In addition to similar test-retest reliability, six Bonferroni-corrected paired-samples *t* tests performed between each baseline showed that there was no significant difference between the mean N1 peak amplitudes for the four baseline windows ($p > .05$).

Because all four baseline windows tested here yielded excellent reliability and the choice of baseline window did

Table 1. Intraclass correlation coefficient (ICC) values for four baseline windows.

Baseline window	<i>N</i>	110 dB pSPL	
		ICC(2, <i>k</i>)	95% confidence interval
Long prestimulus [−11.6, −0.6]	44	0.806	0.645, 0.894
Long whole-trial [−11.6, +10.4]	44	0.809	0.651, 0.896
Short prestimulus [−2.6, −0.6]	43	0.763	0.563, 0.872
Short whole-trial [−2.6, +10.4]	44	0.813	0.658, 0.898

Note. ICC values between 0.60 and 0.74 indicate good test-retest reliability, and values greater than or equal to 0.75 indicate excellent test-retest reliability.

not affect N1 amplitudes, the short prestimulus baseline window was chosen for further analyses. A prestimulus baseline was chosen because it avoids poststimulus response characteristics, and the short window length was chosen because this closely approximates the recommended duration of a baseline window that is 20% of the total epoch (Luck, 2014).

Peak-to-Peak Versus Peak-to-Baseline Measurements

An N1 peak was present in both 110 dB pSPL blocks for all 44 participants. An N1 was present in both 80 dB pSPL blocks for 42 of the 44 participants and in only one of the two 80 dB pSPL blocks for two older participants. These two participants were excluded from further analysis at 80 dB pSPL. One older participant was excluded from the peak-to-baseline condition because the peak amplitude for one of the two presentations fell above 0 μ V. The P1 was present in all blocks at both levels for all participants.

N1 Peak Amplitudes

A paired-samples *t* test indicated that peak-to-peak N1 amplitudes were significantly larger than baseline-corrected N1 peak amplitudes, $t(172) = -14.16$, $p < .001$. This difference is not surprising given that the different measurement techniques use different references when measuring the N1 peak, as discussed above and illustrated in Figure 1.

Test-Retest Reliability

Test-retest reliability was assessed using the ICC. Figure 3 shows associations between N1 peak amplitudes for each participant across Block Presentations 1 and 2 for measurement type and level. Initial repeated-measures analyses of variance were performed on the N1 amplitudes of each measurement type separately, using presentation number as a within-subjects factor to examine systematic error between the first and second presentations of each stimulus intensity (Weir, 2005). There was no significant difference between the N1 amplitudes of the two presentations for either peak-to-peak measurements [110 dB pSPL: $F(1, 43) = 0.103$, $p = .750$; 80 dB pSPL: $F(1, 41) = 3.255$,

Figure 3. Scatter plot of N1 amplitudes in response to clicks presented at 110 dB pSPL (left) and 80 dB pSPL (right), measured using two common methods of quantifying peak amplitude. Peak-to-peak N1 measurements (circles) were taken from the N1 to the P1 at two time points separated by at least 7 min. Baseline-corrected measurements (triangles) were taken as absolute voltage relative to a 2-ms prestimulus baseline from -2.6 to -0.6 ms. Open symbols indicate older participants; filled symbols indicate younger participants. Lines indicate linear regressions for peak-to-peak measurements (dashed) and baseline-corrected measurements (solid). Diagonal dotted gray line is a reference line indicating a correlation of 1 between Presentations 1 and 2.

$p = .079$] or baseline-corrected measurements [110 dB pSPL: $F(1, 43) = 0.075$, $p = .785$; 80 dB pSPL: $F(1, 41) = 0.381$, $p = .540$], indicating no significant systematic error in N1 measurements between the first and second presentations.

The ICC(2, k) values and 95% confidence intervals for each measurement type at two stimulus levels are shown in Table 2. ICC values for both measurement types at both stimulus levels indicated good-to-excellent test-retest reliability (Cicchetti, 1994). ICC values were compared via z tests (Fisher, 1936) and showed no significant differences between the ICC values of the two measurement types for either click level ($p > .05$).

Stimulus and Participant Predictors

Separate linear mixed-effects models were employed to test the extent to which variability in N1 peak amplitudes is uniquely predicted by age group and click level. Separate models were fit for each N1 measurement technique (peak-to-peak and baseline-corrected).

For peak-to-peak N1 amplitude, an intercept-only model with participant as a random factor and N1 amplitude

as the dependent variable (Model P0) was significantly improved by adding age group as a fixed factor, $\chi^2(1) = 4.671$, $p = .031$ (Model P1). Model fit was further improved by adding click level to Model P1 (Model P2), $\chi^2(1) = 7.061$, $p = .008$. Both age group and click level were found to be significant predictors of peak-to-peak N1 amplitude. Younger participants were found to have larger N1 amplitudes ($M = 0.735 \mu\text{V}$) than older participants ($M = 0.535 \mu\text{V}$), $\beta = 0.233$ ($SE = 0.107$), $t(39) = 2.169$, $p = .036$ (see Figure 4). N1 amplitudes were also found to be larger for the 110 dB pSPL click ($M = 0.700 \mu\text{V}$) than for the 80 dB pSPL click ($M = 0.589 \mu\text{V}$), $\beta = 0.166$ ($SE = 0.062$), $t(122) = 2.685$, $p = .008$, consistent with the known spread of excitation in the cochlea and increase in neural synchrony as stimulus intensity increases. Adding the interaction term for age and click level failed to improve model fit, $\chi^2(1) = 0.129$, $p = .720$.

For baseline-corrected N1 amplitude, an intercept-only model with participant as a random factor and N1 amplitude as the dependent variable (Model B0) was significantly improved by adding age group as a fixed factor,

Table 2. Intraclass correlation coefficient (ICC) values and 95% confidence intervals for two measurement types and two stimulus levels.

Condition	<i>N</i>	ICC(2, <i>k</i>)	95% confidence interval
110 dB pSPL			
Peak-to-peak (N1–P1)	44	0.843	0.712, 0.914
Baseline-corrected (2-ms prestimulus)	43	0.763	0.563, 0.872
80 dB pSPL			
Peak-to-peak (N1–P1)	42	0.781	0.475, 0.848
Baseline-corrected (2-ms prestimulus)	42	0.678	0.401, 0.827

Figure 4. N1 amplitudes quantified by peak-to-peak measurements from the N1 to the P1 (left) and by baseline-corrected measurements (right) in response to N1 clicks presented at 110 and 80 dB pSPL. Baseline-corrected measurements were taken as absolute voltage relative to a 2-ms prestimulus baseline from –2.6 to –0.6 ms. Horizontal black bars indicate the average across participants for each condition. Filled triangles indicate younger participants; open circles indicate older participants. Each participant heard each click twice, yielding two data points per participant.

$\chi^2(1) = 3.880$, $p = .049$ (Model B1). Model fit was further improved by adding click level to Model B1 (Model B2), $\chi^2(1) = 11.071$, $p = .001$. Similar to the models evaluating peak-to-peak N1 amplitude, both age and click level were found to be significant predictors of baseline-corrected N1 amplitude. Younger participants were found to have larger N1 amplitudes ($M = 0.461 \mu\text{V}$) than older participants ($M = 0.343 \mu\text{V}$), $\beta = 0.233$ ($SE = 0.118$), $t(39) = 1.967$, $p = .056$ (see Figure 4). N1 amplitudes were also found to be larger for the 110 dB pSPL click ($M = 0.449 \mu\text{V}$) than for the 80 dB pSPL click ($M = 0.368 \mu\text{V}$), $\beta = 0.186$ ($SE = 0.055$), $t(122) = 3.390$, $p < .001$. Adding the interaction term for age and click level failed to improve model fit, $\chi^2(1) = 1.098$, $p = .295$.

To examine the extent to which age and click level predictions of N1 amplitude were affected by measurement types, the standardized regression coefficients for age and click level from Models P2 and B2 were compared using Z tests (see Table 3). The standardized estimates for age and click level did not significantly differ between Models

P2 and B2 (see Table 3), suggesting that these two factors predict AN function similarly when AN function is measured as peak-to-peak and baseline-corrected N1 amplitudes.

Discussion

In this study, we assessed the extent to which measurement type affects test–retest reliability of N1 peak amplitude measurements of the CAP. We showed that both peak-to-peak and baseline-corrected measures result in N1 responses with good-to-excellent test–retest reliability and that peak-to-baseline measures yield reliable responses regardless of the choice of baseline window. These new findings are promising and suggest that CAPs can be measured reliably, regardless of measurement type or baseline window. We also assessed how age and click level affect N1 amplitude and examined how measurement type affects these relationships. The results suggest that age and click level similarly predict N1 amplitude when using peak-to-peak and baseline-corrected techniques. Therefore, despite

Table 3. Standardized estimates of linear mixed-effects models.

Term	Model P2			Model B2			Comparison	
	β	SE	n	β	SE	n	$Z_{\text{difference}}$	$p_{\text{two-tailed}}$
Age group	–0.233	0.107	41	–0.233	0.118	41	< 0.001	1.000
Click intensity level	0.166	0.062	41	0.186	0.055	41	0.241	.809

the expected absolute differences in amplitudes between peak-to-peak and peak-to-baseline measurements, both measures similarly characterize differences between age groups and the effects of click level. Benefits of peak-to-peak measurements include larger responses and higher clinical availability. However, these measures require the identification of two peaks rather than one, and resulting measurements are dependent on factors that may affect both N1 and P1. Benefits of peak-to-baseline measurements include decreased computational demand, greater specificity, and consistency with EEG methodologies.

Defining the peaks surrounding N1 can be difficult in certain ECoG contexts, particularly for participants with diminished responses, as was the case for some of our older participants and for investigators less experienced with identifying peaks. In these situations, proper P1 peak identification can present a serious challenge to accurate measurements of the N1. In such situations, peak-to-baseline measurements may provide a suitable and computationally simpler alternative.

When peak-to-baseline measurements are used, the results of our study suggest that the selection of baseline window has little effect on the outcome. However, this may be because of the preprocessing we performed as part of our data analysis, which included high-pass filtering and artifact rejection. In the absence of high-pass filtering, we would anticipate that the choice of baseline window would have a greater effect (Luck, 2014; Tanner et al., 2016). Regardless, because the choice of baseline window had minimal effect on N1 amplitudes or their test–retest reliability, we suggest choosing a baseline that is appropriate for the given stimuli and apparatus. A prestimulus baseline is typically selected, so that the baseline window does not overlap with the neural response, and many investigators recommend a baseline window that ends shortly before stimulus presentation ($t = 0$) to avoid temporal smoothing and the presence of either stimulus artifacts or neural responses in the baseline (M. X. Cohen, 2014). A baseline period that is at least 20% of the total epoch duration is also recommended (Luck, 2014).

We were able to reliably measure CAP N1 amplitudes in a large cohort of older and younger participants. We found that both peak-to-baseline and peak-to-peak measurements have good-to-excellent reliability for quantifying the amplitude of the N1 peak of the CAP in humans. We also found that measurement type did not affect the accuracy with which N1 amplitude was predicted by experimental factors such as participant age and click intensity level. Similar to previous studies, N1 amplitudes were larger for younger participants than older participants (Konrad-Martin et al., 2012) and were larger for higher click levels than lower click levels (Harris et al., 2018; Thornton, 1976), regardless of the way N1 amplitude was quantified. Taken together, these findings show that both are reliable methods for quantifying N1 amplitude and enable clinicians and researchers to compare findings across studies that use either of these two measurement types.

Acknowledgments

This work was supported (in part) by grants from the National Institute on Deafness and Other Communication Disorders (R01 DC014467 [Dr. Harris], P50 DC00422 [Dr. Dubno], and T32 DC014435 [Dr. Dubno]). The project also received support from the South Carolina Clinical and Translational Research Institute with an academic home at the Medical University of South Carolina, National Institutes of Health/The National Center for Research Resources Grant UL1RR029882. This investigation was conducted in a facility constructed with support from Research Facilities Improvement Program Grant C06 RR14516 from the National Center for Research Resources, National Institutes of Health. We thank the participants of our study.

References

- Baba, A., Takasaki, K., Tanaka, F., Tsukasaki, N., Kumagami, H., & Takahashi, H. (2009). Amplitude and area ratios of summing potential/action potential (SP/AP) in Ménière's disease. *Acta Oto-Laryngologica*, 129(1), 25–29. <https://doi.org/10.1080/00016480701724888>
- Bourien, J., Tang, Y., Batrel, C., Huet, A., Lenoir, M., Ladrech, S., ... Wang, J. (2014). Contribution of auditory nerve fibres to compound action potential of the auditory nerve. *Journal of Neurophysiology*, 112, 1025–1039. <https://doi.org/10.1152/jn.00738.2013>
- Bramhall, N. F., Konrad-Martin, D., McMillan, G. P., & Griest, S. E. (2017). Auditory brainstem response altered in humans with noise exposure despite normal outer hair cell function. *Ear and Hearing*, 38, e1–e12. <https://doi.org/10.1097/AUD.0000000000000370>
- Chertoff, M. E., Lichtenhan, J. T., & Willis, M. (2010). Click- and chirp-evoked human compound action potentials. *The Journal of the Acoustical Society of America*, 127, 2992–2996. <https://doi.org/10.1121/1.3372756>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6, 284–290.
- Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, 100, 1261–1293. <https://doi.org/10.1086/230638>
- Cohen, J. E., Cohen, P., West, S. G., & Aiken, L. S., (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cohen, M. X. (2014). *Analyzing neural time series*. Cambridge, MA: MIT Press.
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *The Journal of Neuroscience Methods*, 134, 9–21. <https://doi.org/10.1016/j.jneumeth.2003.10.009>
- Earl, B. R., & Chertoff, M. E. (2010). Predicting auditory nerve survival using the compound action potential. *Ear and Hearing*, 31, 7–21. <https://doi.org/10.1097/AUD.0b013e3181ba748c>
- Eggermont, J. J. (1976a). Analysis of compound action potential responses to tone bursts in the human and guinea pig cochlea. *The Journal of the Acoustical Society of America*, 60, 1132–1139. <https://doi.org/10.1121/1.381214>
- Eggermont, J. J. (1976b). Summating potentials in electrocochleography: Relation to hearing disorders. In R. J. Ruben, C. Elberling, & G. Salomon (Eds.), *Electrocochleography* (pp. 67–87). Baltimore, MD: University Park Press.

- Eggermont, J. J.** (2017). Ups and downs in 75 years of electrocochleography. *Frontiers in Systems Neuroscience*, *11*, 1–21. <https://doi.org/10.3389/fnsys.2017.00002>
- Fisher, R. A.** (1936). Statistical methods for research workers. *Biological Monograph Manuals*, *5*, 360.
- Gibson, W. P., Moffat, D. A., & Ramsden, R. T.** (1977). Clinical electrocochleography in the diagnosis and management of Ménière's disorder. *Audiology*, *16*, 389–401. <https://doi.org/10.3109/00206097709071852>
- Hallgren, K. A.** (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*, 23–34. <https://doi.org/10.20982/tqmp.08.1.p023>
- Harris, K. C., Vaden, K. I., Jr., McClaskey, C. M., Dias, J. W., & Dubno, J. R.** (2018). Complementary metrics of human auditory nerve function derived from compound action potentials. *Journal of Neurophysiology*, *119*, 1019–1028. <https://doi.org/10.1152/jn.00638.2017>
- Hu, L., Xiao, P., Zhang, Z. G., Mouraux, A., & Iannetti, G. D.** (2014). Single-trial time-frequency analysis of electrocortical signals: Baseline correction and beyond. *NeuroImage*, *84*, 876–887. <https://doi.org/10.1016/j.neuroimage.2013.09.055>
- Konrad-Martin, D., Dille, M. F., McMillan, G. P., McDermott, D., Fausti, S. A., & Austin, D. F.** (2012). Age-related changes in the auditory brainstem response. *Journal of the American Academy of Audiology*, *35*, 18–35.
- Koo, T. K., & Li, M. Y.** (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kujawa, S. G., & Liberman, M. C.** (2009). Adding insult to injury: Cochlear nerve degeneration after “temporary” noise-induced hearing loss. *The Journal of Neuroscience*, *29*, 14077–14085. <https://doi.org/10.1523/JNEUROSCI.2845-09.2009>
- Liberman, M. C., Epstein, M. J., Cleveland, S. S., Wang, H., & Maison, S. F.** (2016). Toward a differential diagnosis of hidden hearing loss in humans. *PLoS One*, *11*, 1–15. <https://doi.org/10.1371/journal.pone.0162726>
- Lichtenhan, J. T., & Chertoff, M. E.** (2008). Temporary hearing loss influences post-stimulus time histogram and single neuron action potential estimates from human compound action potentials. *The Journal of the Acoustical Society of America*, *123*, 2200–2212. <https://doi.org/10.1121/1.2885748>
- Lopez-Calderon, J., & Luck, S. J.** (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*, 1–14. <https://doi.org/10.3389/fnhum.2014.00213>
- Luck, S. J.** (2014). *An introduction to the event-related potential technique* (2nd ed.). Cambridge, MA: MIT Press.
- McGraw, K. O., & Wong, S. P.** (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Mehraei, G., Gallardo, A. P., Shinn-Cunningham, B. G., & Dau, T.** (2017). Auditory brainstem response latency in forward masking, a marker of sensory deficits in listeners with normal hearing thresholds. *Hearing Research*, *346*, 34–44. <https://doi.org/10.1016/j.heares.2017.01.016>
- Mehraei, G., Hickox, A. E., Bharadwaj, H. M., Goldberg, H., Verhulst, S., Liberman, M. C., & Shinn-Cunningham, B. G.** (2016). Auditory brainstem response latency in noise as a marker of cochlear synaptopathy. *The Journal of Neuroscience*, *36*, 3755–3764. <https://doi.org/10.1523/JNEUROSCI.4460-15.2016>
- Murnane, O. D., Prieve, B. A., & Relkin, E. M.** (1998). Recovery of the human compound action potential following prior stimulation. *Hearing Research*, *124*, 182–189. [https://doi.org/10.1016/S0378-5955\(98\)00136-1](https://doi.org/10.1016/S0378-5955(98)00136-1)
- Parthasarathy, A., Datta, J., Torres, J. A. L., Hopkins, C., & Bartlett, E. L.** (2014). Age-related changes in the relationship between auditory brainstem responses and envelope-following responses. *Journal of the Association for Research in Otolaryngology*, *15*, 649–661. <https://doi.org/10.1007/s10162-014-0460-1>
- Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A.** (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*, 859–866. <https://doi.org/10.1111/j.1745-9125.1998.tb01268.x>
- Plack, C. J., Legér, A., Prendegast, G., Kluk, K., Guest, H., & Munro, K. J.** (2016). Toward a diagnostic test for hidden hearing loss. *Trends in Hearing*, *20*, 1–9. <https://doi.org/10.1177/2331216516657466>
- Santarelli, R., & Arslan, E.** (2002). Electrocochleography in auditory neuropathy. *Hearing Research*, *170*, 32–47. [https://doi.org/10.1016/S0378-5955\(02\)00450-1](https://doi.org/10.1016/S0378-5955(02)00450-1)
- Santarelli, R., Starr, A., Michalewski, H. J., & Arslan, E.** (2008). Neural and receptor cochlear potentials obtained by transtympanic electrocochleography in auditory neuropathy. *Clinical Neurophysiology*, *119*, 1028–1041. <https://doi.org/10.1016/j.clinph.2008.01.018>
- Schmiedt, R. A., Mills, J. H., & Boettcher, F. A.** (1996). Age-related loss of activity of auditory-nerve fibers. *Journal of Neurophysiology*, *76*, 2799–2803.
- Schvartz-Leyzac, K. C., & Pfingst, B. E.** (2016). Across-site patterns of electrically evoked compound action potential amplitude-growth functions in multichannel cochlear implant recipients and the effects of the interphase gap. *Hearing Research*, *341*, 50–65. <https://doi.org/10.1016/j.heares.2016.08.002>
- Scott, W. C., Giardina, C. K., Pappa, A. K., Fontenot, T. E., Anderson, M. L., Dillon, M. . . . Fitzpatrick, D. C.** (2016). The compound action potential in subjects receiving a cochlear implant. *Otology & Neurotology*, *37*, 1654–1661. <https://doi.org/10.1097/MAO.0000000000001224>
- Shrout, P. E., & Fleiss, J. L.** (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.
- Sijtsma, K.** (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*, 107–120. <https://doi.org/10.1007/s11336-008-9101-0>
- Tanner, D., Norton, J. J. S., Morgan-Short, K., & Luck, S. J.** (2016). On high-pass filter artifacts (they're real) and baseline correction (it's a good idea) in ERP/ERMF analysis. *Journal of Neuroscience Methods*, *266*, 166–170. <https://doi.org/10.1016/j.jneumeth.2016.01.002>
- Thornton, A. R. D.** (1976). Statistical properties of electrocochleographic responses and their use in clinical diagnosis. In R. J. Ruben, C. Elberling, & G. Salomon (Eds.), *Electrocochleography* (pp. 257–276). Baltimore, MD: University Park Press.
- Weir, J. P.** (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research*, *19*, 231–240. <https://doi.org/10.1519/15184.1>