DIAGNOSTIC AUDITORY BRAINSTEM RESPONSE ANALYSIS: EVALUATION OF SIGNAL-TO-NOISE RATIO CRITERIA USING SIGNAL DETECTION THEORY

by

RONETTE HABOOSHEH

B.A., University of British Columbia, 2001

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Audiology and Speech Sciences)

THE UNIVERSITY OF BRITISH COLUMBIA

June 2007

Abstract

This study evaluated an online measure of signal-to-noise ratio (SNR) as a response-detection tool for threshold auditory brainstem response (ABR) testing. Threshold-ABR data were analysed for 98 infants and young children tested at BC Children's Hospital, and results were validated on an additional 10 patients. Using signal detection theory, it was possible to assess test performance for the SNR measure, with expert-clinician judgement as the gold standard. In addition, a range of SNR criteria were assessed in terms of sensitivity (the ability to accurately identify a response) and specificity (the ability to accurately reject waveforms that do not contain a response). The effect of residual noise (RN) exclusion criteria on SNR test performance, sensitivity, and specificity was also investigated. Waveforms to 500-, 2000-, and 4000-Hz air-conducted brief-tone stimuli were included in this study. Overall, SNR was found to have a test performance of A=.91, with improved performance (A=.93) when high residual-noise waveforms (RN>0.08 μV) were excluded. When low-RN data were separated by frequency, test performance for each frequency was A=.94. Results suggest that the optimal SNR criterion is slightly lower for 500-Hz recordings than for 2000- or 4000-Hz recordings. However, when high-RN recordings were excluded, a SNR criterion of 0.98 achieved a minimum specificity of 95% for each stimulus frequency, with sensitivity values ranging from 64%(for 500 Hz) to 79% (for 4000 Hz). Findings confirm the hypotheses that SNR accurately distinguishes response-present from response-absent waveform, and that quiet recordings are more easily interpreted than noisy recordings using SNR. Guidelines are provided for the clinical use of SNR as an objective response-detection tool.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
ABR	Auditory brainstem response
BCCH	British Columbia's Children's Hospital
BCEHP	British Columbia Early Hearing Program
CCR	Correlation coefficient across replications
EEG	Electroencephalogram
EHL	Estimated behavioural hearing level
IHS	Intelligent Hearing Systems
N	Number of subjects
nHL	Normal hearing level
OIHP	Ontario Infant Hearing Program
POVR	Point optimized variance ratio
ppe	Peak-to-peak equivalent
RN	Residual noise
ROC	Receiver operating characteristic
SD	Standard deviation
SDR	Standard deviation ratio
SL	Sensation Level
SNR	Signal-to-noise ratio (IHS Smart-EP measure)
SPL	Sound pressure level

Acknowledgements

Many thanks to my supervisor, Dr. David Stapells, for the time and effort he devoted to my thesis. His guidance and expertise have been invaluable to me. I would also like to thank my committee members, Dr. Navid Shahnaz and Laurie Usher, for their insight and kind support, as well as the expert judges, and everybody at the BCCH Audiology Department without whom this project would not have been possible. Last, but not least, I am extremely grateful to my family and friends for their patience, generosity and encouragement.

This research was generously supported by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant awarded to Dr. Stapells.

Literature Review

The auditory brainstem response (ABR) is a short latency (<20 ms post-stimulus) recording of the EEG in response to an auditory stimulus. ABRs to brief tones can be used to estimate hearing thresholds, yielding data on type, degree and configuration of loss in patients who cannot be adequately tested using behavioural techniques (Beattie, 1998; Picton, Durieux-Smith, & Moran, 1994; Stapells, 2000a; Stapells, 2000b). Reliable electrophysiological data can be acquired at a younger age than behavioural thresholds (Hyde, Riko, & Malizia, 1990; Stapells, 2000a), for normal as well as developmentally-delayed patients. As a result, tone-evoked ABR is an integral component of programs for the early identification of hearing loss. Early identification allows for earlier intervention and leads to better outcomes for children with hearing impairment. ABR findings inform intervention planning (including cochlear implant candidacy), and assist the community audiologist in selecting and fitting amplification. ABR results must be as accurate as possible in order to maximize each child's potential for hearing, while maintaining comfort and avoiding damage due to overamplification. Despite the need for high levels of accuracy and precision in ABR testing, clinical protocols continue to emphasize subjective (i.e., visual) response interpretation techniques.

Description of the ABR

The ABR contains potentials generated at the levels of the eighth nerve and brainstem. The first human ABRs were replicated in 1970 by Jewett and colleagues (Jewett, Romano, & Williston, 1970). The response consists of up to five waves, which are identified according to their latency and morphology. Waves I and II reflect the activity of the auditory nerve, while Wave V is generated primarily by pathways leading from the lateral lemniscus to the inferior colliculus of each side (Møeller, Jho, Yokota, & Jannetta, 1995). At intensities near threshold, there are significant changes to ABR morphology. The latencies of the component waves increase, and the waves themselves become smaller in amplitude. Wave V typically occurs between 6 and 15 ms after the stimulus, and is usually the most robust of the waves (Picton, 1990).

ABRs can be elicited by clicks as well as brief tones. ABRs to wideband (click) stimuli are a non-invasive means of obtaining information about the integrity of the eighth nerve and auditory brainstem pathways. In addition, click-evoked ABRs are useful for infant hearing screening programs. As a tool for diagnosing hearing loss, however, the click-ABR is inadequate. Clicks have a broad frequency spectrum, and therefore stimulate a large region on the basilar membrane. Responses from high- and low-frequency regions cannot be easily distinguished (e.g., Don & Eggermont, 1978; Picton, Stapells & Campbell, 1981). Moreover, research has shown that thresholds obtained via click-ABR typically correspond to the lowest (best) behavioural threshold

between 1000 and 4000 Hz (Durieux-Smith, Picton, Bernard, MacMurray, & Goodman, 1991). Consequently, the degree of hearing loss is often underestimated, and the hearing loss configuration cannot be accurately defined (Stapells, 2000a; Oates & Stapells, 1997a).

Current guidelines recommend the use of brief-tone ABR for follow-up of infants who have been referred from a newborn hearing screening (American Academy of Audiology, 2003; American Speech-Language-Hearing Association, 2004; Joint Committee on Infant Hearing, 2000). The use of frequency-specific stimuli greatly enhances our ability to predict behavioural hearing thresholds for each frequency (Stapells, 2000a). The brief-tone stimuli used in ABR have a main lobe of energy centred around the nominal frequency, as well as low-amplitude side lobes at higher and lower frequencies. Furthermore, studies have confirmed that the use of brief-tone stimuli results in frequency-specific thresholds (Oates & Stapells, 1997a, 1997b; Stapells, 2000b). Stapells (2000b) performed a meta-analysis of studies investigating threshold estimation with tone-evoked ABR. The combined subject pool included 1,203 participants with either normal hearing or sensorineural hearing loss. 679 of the participants were infants or young children. Results indicate that for infants and children with sensorineural hearing loss, tone-evoked ABR thresholds are, on average, 5.5 dB above behavioural thresholds at 500 Hz, 0.6 dB above behavioural thresholds at 2000 Hz, and 8.1 dB below (better than) behavioural thresholds at 4000 Hz. Mean

brief-tone ABR thresholds for infants and young children with normal hearing range from 13.6 dB nHL at 2000 Hz to 19.6 dB nHL at 500 Hz. In addition, 95% confidence intervals for all frequencies were within 4 dB of the mean difference scores. These findings indicate that brief-tone ABR provides accurate and reliable estimates of psychoacoustic thresholds.

Whereas responses to moderate-to-high intensity clicks consist of multiple peaks (waves I-V), the most prominent feature of brief-tone ABR is the large slow-wave negativity, which immediately follows wave V (particularly for low-frequency brief tones). This negativity is therefore very important for determining response presence during frequency-specific threshold testing (Takagi, Suzuki, & Kobayashi, 1985). Moreover, responses to tonal stimuli occur at later latencies than those to clicks. Morphological differences also exist between responses to high- and low-frequency stimuli. For example, responses to 500-Hz brief-tones are broader and later in latency than those acquired to 2000- and 4000-Hz stimuli of the same intensity (Stapells, 2000a). Wave V amplitudes, however, are relatively stable across frequencies (Stapells, 2000a).

The relationship between behavioural and ABR thresholds varies with methodological parameters (e.g., Elberling & Don, 1987a), as well as subject factors such as age (Picton, Durieux-Smith, & Moran, 1994). The infant ABR is smaller, morphologically simpler, and shows later peak latencies (Durieux-Smith, Edwards,

Picton, & MacMurray, 1985; Hecox & Galambos, 1974, Picton et al., 1994). The amplitude of wave V is especially reduced in infants compared to adults.

ABR and averaging

The ABR is a far-field measurement (electrodes are placed on the scalp). As a result, it is difficult to distinguish its signals from electrical activity generated by non-auditory parts of the brain as well as other organs (Elberling & Don, 1987a; Picton, Linden, Hamel, & Maru, 1983). ABR amplitudes can be smaller than 50 nV at threshold (Sininger & Masuda, 1990). Fluctuations in background noise, on the other hand can be as large as 20 μV (Sininger, 1993). Averaging is one technique which is commonly employed to overcome the poor signal-to-noise ratio of the ABR. In fact, it is with the advent of averaging machines that ABR detection first became possible (Jewett, 1994). The neural response to an auditory stimulus is essentially stable in amplitude, latency, and morphology over a large number of trials. Like all neural activity, there is some adaptation of the response; however, this adaptation is complete after the first few trials and does not significantly undermine the stability of the response (Don, Allen, & Starr, 1977). Background noise, on the other hand, is theoretically random and is normally independent of the stimulus. Under most conditions the signal remains constant with averaging, whereas the background noise decreases in amplitude as the number of averaged trials grows. Successive trials can therefore be averaged in order to extract the stimulus-synchronous activity which reflects audition from the spontaneous firing of

auditory pathways, and the non-synchronous activity from other sources. Several thousand responses must often be averaged before ABRs can be accurately assessed (Picton et al., 1983).

Visual response classification

The Auditory Brainstem Response is an objective measure of auditory sensitivity in that the patient is not required to actively participate in the testing process. However, it remains a subjective test insofar as waveforms are analysed by potentially biased clinicians. Typically, a clinician will acquire two waveforms, each of which may consist of 2000 stimulus-trials. The clinician will then compare the two waveforms for the presence and replicability of a response at an appropriate latency relative to the stimulus onset. Diagnostic ABR testing follows a decision-tree model. When a recording is judged "response absent", the clinician may decide to increase the stimulus intensity, switch to the other ear, or move on to bone-conduction testing (Stapells, 2000a). If the recording is deemed "response present", the clinician will likely move on to a different frequency. As a result, errors in waveform analysis affect the course of the entire test. Research has shown that response-detection accuracy varies across clinicians depending on their level of experience (Valdes-Sosa, Bobes, Perez-Abalo, Perera, Carballo, & Valdes-Sosa, 1987) and the amount of residual noise in the recording (Don & Elberling, 1996; Valdes-Sosa et al., 1987).

Based on extensive experience with ABR testing, several researchers have proposed guidelines to assist clinicians with the response-detection task. In order to

demonstrate the presence of a response to the stimulus, the clinician should verify that wave V is repeatable over its entire duration (Stapells, 2000a). Moreover, its peak-to-peak amplitude should be a minimum of three times the average difference between replications (Don & Elberling, 1996; Picton et al., 1983). To allow a conclusion that a response is absent, replications should be essentially flat with little or no variation between them, yielding a low residual noise (RN) value. The presence of excessive residual electrical noise may otherwise obscure a true response.

Background noise decreases as additional trials are obtained according to the square root of the number of trials in the average waveform (Hyde, Sininger, & Don, 1998; Picton et al., 1983). Theoretically then, averaging could continue until the RN level is sufficiently small to allow complete confidence in a visual interpretation of the waveform. In the interest of time efficiency, however, clinicians must often weigh the value of obtaining additional trials with the current stimulus, against the need for information at other frequencies or intensities. This is particularly true when testing infants who may wake at any moment, putting an end to the test session.

In order to help clinicians decide when to stop the averaging process for a specific stimulus, Don and Elberling (1996) proposed the following guidelines. The clinician may stop acquiring when: (a) a normal response should be detectable based on the stimulus intensity and the RN level of the recording; (b) it is possible to identify a response whose amplitude (after correction for RN) is equal to or greater than a desired value; (c) sufficient averaging cannot be attained within the available time; or (d) the RN level is brought below a predetermined criterion; or (e) a criterion is reached on a quantitative detection measure (e.g., for signal-to-noise ratio).

The proposed guidelines help to improve reliability and provide support to new clinicians faced with the challenging task of visual waveform analysis. Any form of visual interpretation, however, is inherently subjective. Despite the difficulties associated with visual waveform analysis, it remains the method most commonly used in diagnostic ABR testing worldwide.

Objective response classification

The subjectivity of ABR testing has been recognized as a liability since the 1970s when researchers first investigated computational response detection measures (e.g., Shepherd & McCarren, 1972; Österhammel, Davis, Wier, & Hirsh, 1973; Weinberg & Cooper, 1972). These early methods compared averaged recordings to a standard template. The degree of correlation between the template and the recording was used to determine whether a response was present. The templates used were based on either theoretical or empirical data. The success of theoretically-derived templates is limited by the inherent variability of response morphology from patient to patient. Alternatively, an empirical template may be derived for each patient, based on data acquired during the test session. Although additional test time is required, the use of an empirically-derived template lowers the risk of error (Wicke, Goff, Wallace, & Allison, 1978).

Early response detection paradigms also differed with respect to the model used to derive a template. The template was either based on a prototypical response, or alternatively, on a model of what the EEG would look like in the absence of a response (i.e., noise alone). In the first case, a waveform would be considered response present

if it correlated highly with the template; using a noise-based standard, on the other hand, a waveform would have to differ significantly from the template to be considered response present. An empirical template is more easily acquired for noise-based standards than for those based on a model of response presence. For a subject with severe hearing loss it may be difficult or impossible to obtain a clear example of a response. Conversely, a template of response absence can be obtained using sub-threshold stimuli or the pre-stimulus EEG from any subject.

Based on these observations, Wicke et al. (1978) developed a statistical detection paradigm for click-ABRs using an empirically-derived template founded on the pre-stimulus EEG. The success of the paradigm was evaluated by comparing results against the ratings of four experienced judges. The investigators found that false-positive error rates could be reduced to between 1% and 4%. They did not, however, report the false-negative error rate (i.e., the number of times a response was present, but went undetected) which corresponded to their chosen criterion.

Experimentation with computerized response-analysis paradigms has continued and several newer models have evolved. These programs and statistical tools can be classified according to the theoretical framework on which they are based. The categories include syntactic pattern recognition, artificial neural networks, and statistical feature detection tools.

Syntactic paradigms. Syntactic programs are modelled after the human reasoning process. Rules of interpretation are created which allow the computer to judge every possible ABR output. Because this requires an exhaustive set of rules,

syntactic peak-picking paradigms are relatively inefficient (Sanchez, Riquenes, & Perez-Abalo, 1995).

Artificial neural networks. Artificial neural networks bypass the need for explicit rules by building on a learning model whereby the network picks up complex patterns from an initial data set. The feed-forward, multilayer network is then capable of generalizing the information to new data (Alpsan & Özdamar, 1992; Sanchez et al., 1995). Although these peak-picking paradigms show promise, they have not been adapted to tone-burst ABR and are not commercially available.

Statistical feature detectors. Several quantitative measures have been proposed to assist clinicians in response interpretation. These measures describe some characteristic or feature of an ABR signal, either in the time or the frequency domain. One class of feature detectors quantifies elements of waveform morphology. The Zero Crossing, for example, represents the number of times that the signal fluctuates across the baseline.

Much study has focused on measures which compute either response reliability or signal-to-noise ratio. Within this class are the standard deviation ratio (SDR), Fsp, and the correlation coefficient between replications (CCR) (Picton et al., 1983; Sanchez, Riquenes, and Perez-Abalo, 1995; Schimmel, Rapin, & Cohen, 1974; Wong & Bickford, 1980), all of which are computed in the time-domain. These measures allow the examiner to ensure the quality of the average (e.g., whether it is quiet enough to stop recording, or to say that there is no response), determine when to stop recording, and specify the confidence level of detection (Don & Elberling, 1996; Durrant & Ferraro,

1999; Stapells, 2000a). Furthermore, the objective values in this class are useful for training purposes as well as for comparisons between clinicians (Stapells, 2000a).

Signal-to-noise ratio measures have also been developed for frequency-domain analysis. For example, the amplitude at the expected response frequency may be compared to the amplitude of neighbouring frequencies. The assumption is that the noise spectrum is relatively flat in the spectral region of the response. However, physiological noise often varies in amplitude across the frequency domain. Even when the response occupies a narrow spectral band, the use of neighbouring frequencies to determine noise power is not recommended (Dobie, 1993). Other frequency-domain measures include the magnitude square coherence, phase coherence, and Hotelling T2.

In his review of the available signal-to-noise ratio measures for the ABR, Dobie (1985) concludes that temporal analysis is best where the response occurs as a sharp peak, and at a specific latency in the time domain. For spectrally narrow responses (such as the steady-state evoked potentials), frequency-domain analyses are optimal.

CCR, Fsp, and SDR

Correlation coefficient between replications. The CCR can be employed as a comparison between two independent averages, or between an average and a template (as in the early models cited above). Hyde et al. (1998) label these classes "replicate cross-correlation", and "template cross-correlation", respectively. Although template cross-correlation has potential as a response identification tool, it has power only insofar as a specific response of interest can be accurately defined (e.g., in terms of

shape, amplitude, and latency). The weakness of template-based procedures lies in the necessity to define a "normal" ABR. ABR morphology varies with frequency and intensity. Moreover, for a given stimulus, ABR waveforms may differ substantially within the normal population and are affected by maturation as well as neurological impairment (Hyde et al., 1998). Even empirically-defined noise-based measures, such as that developed by Wicke and colleagues (1978) may be vulnerable to the effects of stimulus artifact and changes in physiologic noise with the addition of the stimulus. The vestibular-evoked myogenic potential is one example of a non-auditory response to the stimulus. This physiologic activity is present only in the post-stimulus interval and may therefore result in the incorrect conclusion that an auditory response is present. In addition, although a pre-stimulus recording may provide the best approximation of noise levels, these levels change over time, and are affected by sleep stage. Template cross-correlation measures are not widely employed, because of the difficulties inherent in selecting an appropriate standard (Hyde et al., 1998).

Replicate cross-correlation (CCR) does not depend on *a priori* knowledge of response morphology; waveshape is emphasized, but only in terms of its reproducibility (Hyde, et al., 1998). The Pearson product-moment correlation coefficient, r, is calculated by comparing EEG amplitudes of the two averages at each point in time. CCR values range from $r = +1$ to $r = -1$ such that two waveforms with identical shape will be perfectly correlated and will have the maximum CCR value ($r = +1$). For uncorrelated waveforms (e.g., in the case where no response is present and the EEG consists of noise alone), r tends to have a value near zero (Hyde et al., 1998). In

addition, the replicate-cross-correlation value can be mathematically (albeit loosely) converted to SDR, yielding an estimate of the signal-to-noise ratio (Picton et al., 1983). Like the other measures of signal-to-noise ratio described in this section, CCR can be evaluated statistically to determine the probability of response presence. CCR can also be calculated as recording progresses, as can SDR and Fsp. If the criterion for any of these measures is reached during a replication (i.e., before the default number of trials have been acquired), the clinician can stop averaging and move on to a new stimulus. Just as noise levels impact SDR and Fsp, CCR values are reduced by excessive noise in the recording. That is, two replications can be poorly correlated because there is no response or because the response is effectively masked by a high level of residual EEG noise. It is also possible for spurious peaks in the noise to replicate, which may lead to the false positive identification of responses (Elberling & Don, 1987a). Unlike SDR and Fsp, CCR measures do not provide a RN estimate.

Signal-to-noise ratio measures differ in terms of the procedure used to estimate signal and noise amplitudes. Time-domain methods typically measure signal amplitude as the amplitude variance of the average over time (note that background noise is included in the average). Noise estimates have been based on a pre-stimulus baseline, no-stimulus trials, variance across trials, or the (±) reference. The use of pre-stimulus baselines and no-stimulus trials are discouraged because of changes in noise over time, as well as the increased test time required to collect the additional data (Dobie, 1985). In order to obtain pre-stimulus baselines, for instance, it is necessary to use a slow presentation rate.

Fsp. Fsp software describes the significance of the signal-to-noise ratio by comparing the variance of all points in the response epoch of the average waveform (as an estimate of signal amplitude) to the variance at a single point between individual sweeps (noise) (Elberling & Don, 1984, 1987a). The presence of a response at the selected time point should not affect the noise estimate insofar as the signal is constant, and any variance is assumed to arise from the noise alone (Hyde et al., 1998; Sininger, 1993). Furthermore, ABR amplitudes are small in comparison to background noise. The contribution of high-noise trials to the average may be decreased in a process called Bayesian weighting (Hyde et al., 1998; Sininger, 1993). Before being entered into the average, each block of 256 sweeps is assigned a weighting factor. High-noise trials receive less statistical weight, whereas low-RN trials are weighted more heavily in the Fsp calculation. The negative effects of artifacts are therefore reduced (Sininger,1993). A standard F-table can be used to determine the probability of response presence in the ABR for a given Fsp value (Elberling & Don, 1984; Hyde et al., 1998).

Point optimized variance ratio. The point optimized variance ratio (POVR) is an adaptation of the Fsp calculation in which the signal amplitude estimate is derived from 10 specific points in the waveform (The ABaer and AOAE Hearing Screening System Users and Service Manual, 2005). These points are chosen according to the expected latency of the response. POVR is currently used with click stimuli in automated infant hearing screening systems (as opposed to diagnostic ABR) (The ABaer and AOAE Hearing Screening System Users and Service Manual, 2005).

Standard deviation ratio. Like Fsp, the standard deviation ratio (SDR) is a means of estimating the signal-to-noise ratio of an ABR waveform. SDR shares with Fsp (and CCR) the advantage of being calculated as recording progresses. SDR and Fsp differ, however, in the methods used to estimate noise amplitudes.

The standard deviation of the average waveform and the standard deviation of the (±) reference comprise the SDR (Picton et al., 1983). The (±) reference is an estimate of the contribution of background noise to the average waveform. It is calculated by alternately adding and subtracting all of the recorded waveforms in a replication, then dividing by the number of trials (Schimmel, 1967). The process of adding one half of the waveforms and subtracting the other half cancels out any consistent response to the stimulus (i.e., the signal). Insofar as the noise is random, it is not cancelled by the alternate addition and subtraction of waveforms. As would be expected, the (±) reference diminishes in amplitude as a greater number of trials are obtained because of the averaging process (Hyde et al., 1998; Picton et al., 1983). Thus, the (±) reference yields a RN estimate which reflects the visible effect of averaging on background noise. Although the validity of this measure is based on the assumption of a constant evoked potential, Hyde et al. (1998) note that this is its only vulnerability and the (±) reference may therefore be the best available estimate of residual noise.

As previously noted, the RN value itself is an important quality indicator. High noise levels indicate that near-threshold responses will be difficult to detect and will require a larger number of trials. At near-threshold intensities, a response may be masked by the EEG noise, therefore the clinician can only conclude that a response is absent when the RN level is reasonably low. This is particularly true for visual response-detection, but RN will also impact the test performance of computational measures.

SDR describes the signal-to-noise ratio by comparing the variance of the average (which includes the signal as well as the noise) to that of the noise component alone (Picton et al., 1983; Picton, Hink, Perez-Abalo, Linden, & Wiens, 1984). SDR therefore provides a gauge of the quality of a recording as well as the significance of a response.

The Intelligent Hearing Systems (IHS, Miami, Fl., U.S.A.) "Smart-EP" diagnostic ABR machine provides online SDR (and RN) calculations. Their measure is labelled SNR (for signal-to-noise ratio) and its values have exactly half the magnitude of those derived from SDR calculations. Although this measure has the potential to objectify ABR response identification, it is currently of limited use to clinicians as appropriate criteria have not been established or assessed. Establishing SNR criteria would be particularly useful as most pediatric audiology clinics in British Columbia use the Smart-EP system for ABR tests in infants.

Statistically versus empirically defined detection criteria

SDR, CCR and Fsp can be described as statistical measures insofar as they quantify and mathematically evaluate features of the ABR waveform. At issue, however, is whether criteria for these measures should be selected using a statistical or an empirical approach. Both methods can be used estimate the likelihood of response presence for a range of values.

In the statistical approach, the probability estimate is based on a hypothetical distribution. Fsp, for example, is a ratio of variances, and its values should theoretically follow a known F-distribution, F (v1, v2) where v1 and v2 represent the degrees of freedom of the noise and signal estimates, respectively (Don, Elberling, & Waring, 1984; Elberling & Don, 1984).

When Wong and Bickford (1980) first proposed a statistical tool for response detection, [also based on the (±) reference], they cautioned that F values can be statistically assessed only where the degrees of freedom are known. Degrees of freedom (*df*) in this case are the number of independent data values in a given estimate of variance. (±) reference-based measures calculate variance over the time-domain of an average, and these successive sample points are highly correlated. The result is a very small *df* in both the numerator and the denominator of the variance ratio, yielding poor statistical precision (Hyde et al., 1998; Picton et al., 1983). The *df* are affected by the frequency spectrum of the noise, as well as the high-pass filter settings (Hyde et al., 1998; Picton et al., 1983; Wong & Bickford, 1980). The formula 2 X bandwidth X analysis-time (Bendat & Piersol, 1971) can be used to estimate the *df* of an average. At

low intensities, where the ABR consists primarily of low-frequency energy, the theoretical *df* are significantly reduced (e.g., two nearby points in a low-frequency sine-wave will tend to have similar amplitude values). Consequently, statistically derived criteria of response significance would be relatively conservative (Picton et al., 1983), particularly in the case of threshold ABR. This limitation affects the selection of SDR, SNR, CCR, and Fsp criteria alike, insofar as sampling occurs in the time-domain.

Fsp partially addresses the problem of point-to-point correlation by calculating the noise estimate across trials rather than across points in the response time window. If consecutive sweeps are independent, and many sweeps are acquired (>250), then the *df* of the denominator will be large (i.e., in this case, 250 *df*) (Elberling & Don, 1984; Hyde et al., 1998). However, 60-Hz line noise is a potential source of correlation across sweeps, and may artificially inflate the estimate of signal amplitude or that of the noise measurement. The variance ratio may be affected either positively or negatively by this 60-Hz contamination (Hyde et al., 1998). Although the Fsp paradigm could be modified to overcome this problem (e.g., by selecting different analysis points), such adaptations have yet to be implemented commercially.

Because the variance of the average is calculated across the time domain, Fsp offers no advantage in terms of *df* in the numerator. Elberling and Don (1984) empirically determined the *df* using no-stimulus trials and selected an estimated lowest value (*df* =5) to determine critical Fsp values. Given *df*s of 5 and 250 in the numerator and denominator respectively, and using the F-table as a reference, the authors conclude that a specificity of 99% can be reached. Hyde et al. (1998) point out that the

true *df* of an average may differ from 5, leading to the selection of an inappropriate criterion.

In light of the difficulties encountered in statistical analysis it seems appropriate to pursue empirically determined cut-off values for response detection. An empirical criterion is one based on experimentally observed results. In this case, an SDR criterion can be established by collecting a large sample of waveforms and plotting the number of times that a response was present at each SDR value. Similarly, the distribution of SDR scores should be plotted for response-absent waveforms. Cut-off points which fall near the intersection of these two distributions can then be evaluated based on their ability to distinguish between response-present and response-absent waveforms.

Appropriate use of the empirical approach requires that the distributions of the sample adequately represent the distribution in the clinical population. This can be easily accomplished by drawing data from a reasonably large clinical sample. Furthermore, it eliminates the need for (potentially inaccurate) assumptions about the shape of the distributions and the interdependence of variables.

Response detection windows

It should be emphasized that none of the signal-to-noise ratio paradigms discussed (SDR/SNR, Fsp, POVR, CCR) are able to distinguish between auditory responses and other time-locked neural events (Sininger, 1993). Consequently, it is important to define the analysis window such that the major components of the ABR are included, but artifacts arising from the stimulus itself or non-auditory potentials are

avoided. The selection of appropriate response-detection windows should be based on observations of a large sample of real ABRs at each stimulus frequency. From these recordings, the latencies of responses and major noise components (e.g., artifact) should be determined.

Signal detection theory and ABR classification

The performance of a response-detection paradigm hinges on its ability to distinguish between response-present and response-absent waveforms. This ability can be quantified using Signal Detection Theory (SDT), which describes the relationship between sensitivity, specificity, and overall test performance. Sensitivity refers to the probability of a positive test outcome when the variable of interest (e.g., disease) is present. Conversely, specificity is the probability of a negative test outcome when that variable is absent (Hyde, Davidson, & Alberti, 1991). SDT allows accuracy to be compared across a number of tests and in a common metric, independent of the specific cut-off point used. A range of cut-off points is chosen for each test, and outputs with a given criterion are classified as hits, misses, false alarms, or correct rejections. For example, in order to measure the accuracy of SNR in response identification, an initial cut-off point of 1.2 might be chosen. Waveforms would then be classified as either "response present" or "response absent" such that all recordings with a SNR value of greater than 1.2 are considered to contain a response and vice versa. The accuracy of the test under this criterion is measured by comparing its results against some gold standard (e.g., expert clinician judgment). Four measures of accuracy are calculated: (a) the number of hits, or correct identifications of a response;

(b) the number of misses where a true response was not identified; (c) the false positive rate, or incorrect identifications of a signal when no response was present; and (d) the rate of correct rejections for waveforms that did not contain a response. Once a suitable range of criteria is evaluated, hit and false positive rates can be plotted to form a receiver operating characteristic (ROC) curve. In instances where both conditions (normal and pathological) show normally distributed test scores, diagnostic accuracy can be compared by calculating the distance between the means of each of these distributions. This difference is labeled *d'*. The larger the *d'* value, the better able a test is to distinguish between two conditions. To clarify, a test in which normal and pathological cases receive overlapping scores (and which therefore has a small *d'*) will not accurately differentiate the two conditions. The overall ability of a test to discriminate normal from pathological cases may also be quantified by calculating the area under the ROC curve (*A*). *A*-value comparisons are possible even when test scores are not normally distributed. Tests with greater values of *A* (and therefore a greater area under the ROC curve) are inherently better than those with smaller values. The *A* value of a test is stable, whereas sensitivity and specificity vary inversely with the operating point. *A* values can range from 0.5 for a useless test to 1.0 for a perfect test. In terms of SNR, an *A* value of 0.90, would indicate that a randomly selected response-absent waveform has a smaller SNR value than that of a randomly selected response-present waveform 90% of the time (Zweig & Campbell, 1993).

When performing ROC analyses of test performance, it is important to select a gold-standard which reflects the truth as closely as possible. In order to evaluate a new

test of hearing loss, for instance, the gold-standard must be able to accurately determine whether or not hearing loss is actually present. Otherwise, a poor *A* value may arise when the new test is, in fact more accurate than the gold-standard (i.e., because of a disagreement between the two tests) (Zweig & Campbell, 1993). For the discrimination of hearing-impaired versus normal-hearing infants using ABR, follow-up behavioural measures are often used as a gold-standard. That is, ABR thresholds can be compared with those acquired from visually-reinforced, or play audiometry. Behavioural testing usually cannot be carried out until the patient is at least 6 months old (older for patients with developmental disabilities), thus validity is affected by changes in hearing status in the interval between the two tests. For a test of response-presence versus response-absence in ABR waveforms, follow-up behavioural measures are an inappropriate gold-standard for an additional reason. A behavioural gold-standard causes a confound between two factors: the ability of the measure to discriminate a signal in the presence of noise, and the overall correspondence between ABR signal presence versus behavioural threshold. In order to avoid confounds, the gold-standard and the candidate test must be measures of the same variable. For ABR response-detection, the current gold-standard is visual detection by an expert in waveform analysis.

In addition to providing data on the accuracy of the test over a range of operating points, ROC curves allow individuals to assess the effect of each criterion on sensitivity and specificity. Examiners can therefore tailor a test to their goals by selecting the appropriate operating point (Hyde et al., 1991). For the purposes of selecting the optimal SNR criterion, the overall accuracy of SNR as a test of response presence is

less critical than the availability of a specific operating point which yields the desired sensitivity and specificity. The criterion of choice can vary depending on the consequences of the various test outcomes (Hyde et al., 1991). In order to define the desired sensitivity and specificity of a test, a quantitative "utility" can be assigned to each possible outcome. Values may be based on the cost of the procedure and therapies, the benefits of early detection, and the social or emotional consequences of a diagnosis. The long-term expected utility of a test result depends on the probability of its occurrence, as well as its inherent utility. For example, in a population where the prevalence of a disease is low, the optimal operating point will have a smaller false positive rate than for a different population with a higher prevalence of the same disease. Examiners can thus select the operating point which yields the greatest long-term value for the population of interest.

In the case of ABR waveform classification, false negative errors (i.e., missing a response that is actually present) are less problematic than false positive errors (i.e., incorrectly stating that the waveform is response present, when there is no physiologic response to the stimulus). False negative errors could lead to over-estimation of a hearing loss, and the potential for over-amplification. Inappropriate amplification can have a negative impact on hearing aid benefit, and can undermine the overall effectiveness of early intervention. However, the decision-tree procedure for ABR testing provides a safeguard against false-negative errors. When a clinician determines that there is no response to a stimulus, the logical next step is to acquire data for the same frequency at a higher intensity. The clinician then has latency and morphology

information from the higher-intensity (i.e., superthreshold) recording that can be used to re-evaluate the lower-intensity waveform. Thus, further testing provides an opportunity for correction of the error.

False-positive errors are less likely to be detected within a test session than false-negative errors. This is because, once a clinician decides that a response is present, testing for that frequency is complete. The clinician's task is to acquire data at all frequencies in both ears, for both air and bone conduction if necessary. There is seldom time to re-check results at a given intensity, or to test higher intensities once a response has been identified. Importantly, the consequences of false-positive errors are potentially worse than those associated with false-negative errors. Hearing loss may be underestimated, or completely unidentified. In the case of underestimated hearing loss, a child may receive inadequate amplification until follow-up testing occurs as much as six months later. An unidentified hearing loss may impact speech-language, academic and social development (Carney & Møeller, 1998; Yoshinaga-Itano, Sedey, Coulter, & Mehl, 1998; Yoshinaga-Itano, 2003). Based on the above considerations, an optimal SNR criterion should aim to have near-perfect specificity (e.g., 95%), with reasonable sensitivity (e.g., 70% or higher).

Comparing response-detection paradigms: Research to date

In a series of experiments (Don, Elberling, & Waring, 1984; Elberling & Don, 1984, 1987a, 1987b), Don, Elberling and colleagues proposed and evaluated their Fsp measure. In their first paper, Elberling and Don (1984) estimated the distribution of

background noise using 80 dBnHL clicks in 10 normal-hearing subjects. Analysis of their results indicated that the single-point variance noise estimate follows a Gaussian distribution. Once this condition was satisfied, they were able to estimate the *df* of physiological background noise from no-stimulus trials in eight subjects. The authors found that the resulting Fsps fit F-distributions corresponding to *df* values between 8 and 22. A *df* of 5 was selected to reflect a worst case scenario. Given *df* of 5 and 250 in the numerator and denominator respectively, the investigators calculated the Fsp values corresponding to each SNR for a given confidence level. An Fsp of 3.1 was selected as the optimum criterion, yielding a statistically-determined false alarm rate of 1%. An estimate of the hit rate was not provided.

As a follow-up to this study, Don, Elberling, and Waring (1984) compared ABR thresholds derived using an Fsp criterion of 3.1 with psychoacoustic thresholds to the click stimulus. Six subjects participated in this study. The authors found that the criterion was exceeded within 10000 sweeps for stimuli between 5 and 20 dB SL. Furthermore, findings indicated that Fsp was more effective at identifying near-threshold responses than the Wong and Bickford method, when criteria were chosen statistically. They noted, however, that the (±) reference-based measure could not be meaningfully evaluated using F-statistics because of the correlation between the numerator (i.e. the variance of the average which includes the signal and noise), and the denominator which is an estimate of the noise alone.

In order to assess the performance of their Fsp measure, Elberling and Don (1987a) recorded click ABRs from an additional 10 normal-hearing subjects, at

intensities of 34 to 52 dB peak-to-peak equivalent (ppe) SPL. Based on their results, the authors calculated that ABR amplitude would have reached zero at approximately 32 dB ppe SPL (2 dB below the lowest test intensity). The median perceptual threshold to the stimulus was 33 dB ppe SPL. The authors found a mean difference of 5 dB between behavioural thresholds and Fsp-based ABR thresholds (criterion = 3.1) after 10000 sweeps. They noted, however, that there would be a greater disparity between ABR and psychophysical thresholds if fewer sweeps were acquired.

Based on the above ABR data, Elberling and Don (1987b) created ROC curves using statistical definitions of response and no-response (based on *df* of 5 and 250). Results confirmed that the Fsp=3.1 criterion yielded a 1% false positive rate, but at the expense of a low (55%) hit rate (at a stimulus level of 38dB ppe SPL). An increase in the quality of the recording (i.e., a higher signal-to-noise ratio) led to improved test performance, and vice versa.

Sininger (1993) refers to an unpublished study in which Fsp-based response criteria are evaluated for an infant screening program. Fsp values were calculated from no-stimulus ABR recordings in newborns to approximate the Fsp distribution of newborns whose thresholds were worse than the screening level. An Fsp distribution was also obtained from newborns with normal hearing, using 30 dBnHL clicks. Almost no overlap was found between the two distributions. A Fsp value of 2.4 was selected as the optimal criterion, yielding a sensitivity of 98.5% and a specificity of 99%. There was a substantial difference between the optimal criterion in the unpublished study (Fsp

= 2.4) versus the cutoff point selected in the earlier studies (Fsp = 3.1). This resulted from differences in the filter settings used which led to an increase in the estimated *df* of the numerator (*df* = 10 versus *df* = 5 in the earlier studies).

Empirical data from a study on 2995 infants (Norton, Gorga, Widen, Folsom, Sininger, Cone-Wesson, Vohr, Mascher, & Fletcher, 2000) suggests somewhat poorer results. The authors compared results from click-ABRs at 30 dBnHL obtained in the neonatal period against visual reinforcement audiometry (VRA) testing at 8 to 12 months of age. The subject pool included both normal-hearing and hearing-impaired infants. Overall, ROC analyses yielded test performance values of *A*= 0.69 when the 3-frequency pure-tone average (1, 2, and 4 kHz) was used to define hearing loss. When instances of progressive hearing loss, and transient hearing loss at behavioural testing were excluded, test performance reached A=.87. Visual analysis of the cumulative distribution figures presented indicate that an Fsp criterion of 2.4 led to the correct identification of hearing loss in approximately 85% of cases, and the incorrect identification of hearing loss in roughly 12% of normal ears. The stricter criterion of Fsp = 3.1 led to approximately 88% correct identifications and 18% incorrect identifications. Because the gold standard was a follow-up behavioural measure, these data are compromised by the differences between neurologic thresholds to click stimuli in infancy and behavioural responses to tonal stimuli up to 12 months later.

Wu and Stapells (2001) presented their research on Fsp response identification for tone-evoked ABR in nine normal-hearing infants. ABRs were evoked using 500- and 2000-Hz brief tones, as well as clicks at 65, 40, 35, and 25 dB nHL. Additionally,

two no-stimulus recordings were obtained for each subject. Fsp values were calculated offline, and a criterion of 2.09 was determined from the distributions of no-stimulus recordings. They found that the median number of trials required to reach criterion at 35 dB nHL was 768 for clicks, 3584 for 2000-Hz brief tones, and 2048 for 500-Hz brief tones. For four of the nine subjects, recordings to 500-Hz brief tones (at 35 dB nHL) did not reach the criterion within 10240 trials. The waveforms were also analysed by two experienced judges and their ratings compared with Fsp results. For 25 dB nHL clicks, 35 dB nHL 2000-Hz brief tones, and 40 dB nHL 500-Hz brief tones, all recordings exceeded the Fsp criterion, in agreement with judges' ratings of response presence. At lower stimulus intensities, however, agreement between Fsp and the judges decreased. Furthermore, the investigators expressed concern at the large number of false-positive results obtained from the no-stimulus recordings. In 8 of 18 cases where no stimulus was presented, the Fsp criterion was reached within 10240 trials. 60-Hz contamination is cited as a possible explanation for the large number of false positives. Although Fsp is a theoretically promising tool, the published data yield conflicting results. While high levels of specificity (99%) are reported for adult ABRs to click stimuli (Elberling & Don, 1987b), the one experiment on brief-tone ABR in infants (Wu & Stapells, 2001) found a false-positive rate of 44% (i.e., specificity of 66%) for no-stimulus recordings using a criterion of 2.09. It is possible that a higher criterion would have yielded better results. There is currently no empirical data comparing Fsp values for threshold-level brief-tone ABR in normal-hearing and hearing-impaired individuals. As a consequence of the lack

of data and inconsistencies in the literature overall, it is difficult to estimate the true effectiveness of Fsp for response classification in diagnostic ABR.

Other equally-promising response-detection measures have been investigated and compared. Picton et al. (1983) evaluated the accuracy of statistically defined criteria for SDR, CCR, and RN. Responses to click stimuli were recorded from 10 normal-hearing subjects at a range of intensities [0 to 60 dB sensation level (SL), and a no-stimulus control condition]. Degrees of freedom in the SDR and CCR calculations were estimated based on the analysis time and the expected response bandwidth for low-intensity stimuli. They found that use of the statistically derived criteria (SDR = 2.2; CCR = 0.53) led to a large number of unidentified responses (e.g., at 10 dB SL, five responses failed both SDR and CCR criteria). Using these criteria, the ABR to clicks varied significantly from the background noise for intensities of 20 dB SL, given 2000 trials. The same waveforms were then evaluated using an empirical SDR criterion, adapted from Wong and Bickford (1980). Although both criteria failed to detect some responses, use of the empirical cutoff was more accurate for at least the 20 dB SL condition. All of the no-stimulus control trials were correctly identified as response absent by both statistically and empirically derived operating points. In addition, Picton and colleagues (1983) proposed that precision could be improved near threshold if the calculations were made specifically in the wave V latency region.

Arnold (1985) compared visual detection with objective measures based on correlation, variance ratio, and multiple pre-post z tests. Eight normal-hearing adults were tested using click ABR at a range of intensities from 10 to 40 dB SL. The variance ratio used in Arnold's study bases the noise estimate on the pre-stimulus average. Four experienced judges were asked to rate waveforms on a 6 point scale ranging from "1" ABR definitely present, to "6" ABR definitely absent. The investigator found that visual judgements were statistically the most sensitive, with correlation, variance ratio, and z-tests following in that order. However, she notes that the practical difference between measures was negligible. She therefore concludes that the objective measures are useful for eliminating observer bias and improving consistency among clinicians.

Other studies comparing the relative accuracy of quantitative feature detectors have also shown promising results (Sanchez et al., 1995; Valdes-Sosa et al., 1987). Valdes-Sosa et al. (1987) compared four response detection measures in their ability to discriminate stimulus from no-stimulus ABR conditions. The measures assessed were visual identification, CCR, SDR and T2R (Hotelling T2, a frequency-domain measure). Using click stimuli of 60 and 30 dB nHL, as well as a no-stimulus control condition, ABRs were recorded and subsequently rated by a panel of clinicians. The observer response menu had five options which varied from certainty that the waveform does not contain a significant response (e.g., in the case of a recording from the control condition) to certainty that a response is present. These 5 rating options were plotted as different decision criteria on a ROC curve. A range of criterion values derived from each of the three feature detectors was also evaluated using ROC curves and Signal Detection

methodology (described below). True and false positive rates for all measures were defined according to stimulus presence in the experimental condition.

The ROC curves for visual response analysis showed frequent fluctuations in the response presence criteria, both across observers and for the same observer on a different day. Those clinicians who had more experience with ABR analysis were the most accurate, and all observers showed better discrimination ability for the higher-level stimuli. High residual noise levels were found to increase detection errors. Comparisons between visual response analysis and the statistical indices were complicated by the fact that the ROC curves had different shapes. In the operating region of P (False Alarm) = 0.10, the two most experienced human observers were more effective than all of the feature detectors. One other observer was more accurate than SDR and CCR, and the other three were less accurate than any of the calculation-based measures. Valdes-Sosa and colleagues concluded from their findings that there is value in both visual and statistical analysis techniques. Human observers are able to integrate information about waveshape, latency, and amplitude changes with computational measures of reliability. With experience, therefore, clinicians can develop greater sensitivity and specificity than the statistical tools alone. Moreover, statistical information can be used to enhance the performance of inexperienced clinicians.

In their clinical guidelines, Don and Elberling (1996) suggest that audiologists stop acquiring once a statistical criterion is reached. The premise is that a signal-to-noise ratio criterion may be reached before the clinician is sufficiently confident of visual results. Such a procedure, however, might limit the clinicians ability to analyse waveform characteristics, thereby affecting diagnostic accuracy. It is perhaps for this

reason that automated ABR has been implemented as a screening tool (e.g., POVR), but has yet to be endorsed for diagnostic testing.

Further research by Sanchez et al. (1995) evaluated the accuracy of individual quantitative measures against the performance achieved using multiple features simultaneously to discern response presence or absence in ABRs to 45 dB nHL clicks. The authors found that isolated feature detectors which are based on reliability and signal-to-noise ratios (such as SDR, the linear similarity coefficient, and the Levenshtein distance) each achieved an area equal to or greater than 0.95 under the ROC curve. They noted, furthermore, that the overall accuracy of these feature detectors was higher than reported in studies by Picton et al. (1983) and Valdes-Sosa et al. (1987). This improvement was attributed in part to the use of a more-intense stimulus. Moreover, a smaller time window, centred on the largest peaks, was used for feature calculations. By highlighting the region of interest, they were able to decrease the effects of various artifacts and lower the false positive rate.

Three important conclusions can be drawn from research on ABR analysis techniques. In the first place, the reliability of visual response identification depends on clinician experience. Furthermore, feature detectors in general have the ability to discriminate response presence and absence, with improved precision if the analysis window can be restricted to the region of Wave V. In addition, the use of empirically defined criteria for feature detectors may be appropriate in cases where statistical calculations underestimate the power of a test. Finally, high residual noise levels lead to an increase in response-detection errors.

As noted above, measures of the signal-to-noise ratio have practical advantages over other analysis techniques. Unlike peak-picking paradigms which replace the human observer, simple feature detectors could be used to inform clinicians in their analyses. This reduces the subjectivity of visual identification while taking advantage of human pattern recognition and deductive capabilities. Furthermore, SDR/SNR and Fsp can be measured on-line, and can therefore guide the clinician in the decision to stop averaging. These measures also allow the clinician to operate with a known sensitivity and specificity.

There are, however, several weaknesses in the research on signal-to-noise ratio measures. These include lack of data on: (a) hearing-impaired subjects, (b) brief-tone ABR, (c) near-threshold responses, and (d) (±) reference-based measures. Without adequate research on hearing-impaired populations, it is not possible to evaluate the true test performance of a measure. Moreover, because of substantial differences between response properties to wideband versus tonal stimuli, conclusions for brief-tone (diagnostic) ABR cannot be extrapolated from studies on click-evoked potentials. In a similar vein, morphological differences between supra-threshold and near-threshold responses make it difficult to draw inferences for threshold ABR based on results from high-intensity stimuli. Threshold-level brief-tone ABR is the gold standard for the diagnosis of hearing loss in infancy, and SNR is widely available on diagnostic ABR machines. It follows, therefore, that the effectiveness of SNR as a response-detection tool should be explored under the specific conditions for which it is intended.

Proposed research study

The objective of the present study is to define appropriate SNR criteria for the evaluation of ABR waveforms to 500-, 2000-, and 4000-Hz air-conducted stimuli from a sample of infants with normal or impaired hearing. The sensitivity and specificity of a range of criteria will be evaluated, using expert clinician judgment as the gold standard. The criterion point with the best performance will be recommended for clinical use.

SNR, in particular, is the focus of this study, as a result of its availability on the IHS: Smart-EP diagnostic ABR machine. Pediatric audiology clinics throughout British Columbia have recently been furnished with the Smart-EP through the BC Early Hearing Program, whose goal is early identification and intervention for hearing impaired infants (BCEHP, 2006). Audiologists across the province will soon be performing diagnostic brief-tone ABR, often for the first time in their careers. Depending on the patient population at their respective clinics, these audiologists may or may not have the opportunity to perform ABRs regularly and thereby maintain their expertise. If it is possible to establish SNR criteria, clinicians will gain an additional tool in ABR decision-making. Response detection could conceivably become more accurate, more objective, and more time-efficient, resulting in more-complete diagnostic results. Moreover, intervention could be planned earlier, and hearing aid fittings would be based on more precise threshold estimates. Audiologists can avoid the pitfalls of either under- or over-amplification, and better outcomes can be expected.

A movement towards the systematic use of the signal-to-noise ratio for response detection could similarly impact clinicians outside of British Columbia, once equipment-specific criteria are established. Furthermore, studies on the relationship

between signal-to-noise ratio and response presence have concentrated exclusively on ABRs to click stimuli. Research is needed to determine whether, and to what extent, measures of signal-to-noise ratio are appropriate in the evaluation of brief-tone ABRs.

The current study employs Signal Detection Theory to assess a range of SNR criteria for the classification of diagnostic brief-tone ABR data. Three hypotheses will be examined:

1. SNR values will accurately differentiate present from absent responses to stimuli of at least some frequencies.
2. Because of differences in Wave V morphology across frequencies, SNR measures may not be equally accurate for all stimuli. ABRs to low-frequency stimuli tend to be broader and more rounded (Stapells, 2000a). Consequently, a variance-based analysis such as SNR may not be as effective for low frequency stimuli as it is at high-frequencies.
3. The accurate identification of an ABR waveform as either response present or response absent may depend on the amount of noise in the recording. It is therefore expected that test performance will improve when high-noise recordings are excluded from the analysis.

Introduction

The auditory brainstem response (ABR) to brief tones is a frequency-specific tool for hearing-threshold estimation. It is especially useful for testing infants and young children who cannot be adequately assessed using behavioural techniques (Beattie, 1998; Picton, Durieux-Smith & Moran, 1994; Stapells, 2000a; Stapells, 2000b). Consequently, brief-tone ABR is integral in the drive towards early intervention for hearing loss. ABR findings inform intervention planning, and assist the community audiologist in selecting and fitting amplification. Despite the importance of accuracy and precision in ABR testing, clinical protocols continue to emphasize subjective (i.e., visual) response-interpretation techniques.

The most prominent feature of brief-tone ABR is wave V and the large slow-wave negativity which immediately follows wave V (particularly for low-frequency brief tones). This negativity is therefore an important index of response presence during frequency-specific threshold testing (Takagi, Suzuki, & Kobayashi, 1985). Although ABR testing is objective in the sense that active patient-participation is not required, it remains subjective insofar as waveforms are analysed by potentially biased clinicians. Research has shown that response-detection accuracy varies across clinicians depending on their level of experience (Valdes-Sosa, Bobes, Perez-Abalo, Perera, Carballo & Valdes-Sosa, 1987) and the amount of residual noise (RN) in the recording (Don & Elberling, 1996; Valdes-Sosa et al., 1987). Because diagnostic ABR testing follows a decision-tree model, errors in waveform analysis affect the course of the entire test.

The RN value is an important quality indicator. High noise levels indicate that near-threshold responses will be difficult to detect and will require a larger number of trials. At near-threshold intensities a response may be masked by the residual electrical noise. Therefore, the clinician can only conclude that a response is absent if the RN level is appropriately low. This is particularly true for visual response-detection, but RN will also impact the test performance of computational measures (Valdes-Sosa et al., 1987).

The subjectivity of ABR testing has been recognized as a liability since the 1970s when researchers first investigated computational response detection measures (e.g., Shepherd & McCarren, 1972; Osterhammel, Davis, Wier & Hirsh, 1973; Weinberg & Cooper, 1972). Since then, several algorithms have been proposed for objective analysis of the ABR. Prominent among these are statistical feature detection tools such as CCR, Fsp, and SDR (Picton, Linden, Hamel & Maru, 1983; Sanchez, Riquenes, and Perez-Abalo, 1995; Schimmel et al., 1974; Wong & Bickford, 1980). These measures allow the examiner to ensure the quality of the average (e.g., whether it is quiet enough to stop recording, or to say that there is no response), determine when to stop recording, and specify the confidence level of detection (Don & Elberling, 1996; Durrant & Ferraro, 1995; Stapells, 2000a). Furthermore, the objective values in this class are useful for training purposes as well as for comparisons between clinicians (Stapells, 2000a).

The correlation coefficient across replications (CCR) is a comparison between two independent averages. Fsp describes the significance of the signal-to-noise ratio by comparing the variance of all points in the response epoch of the average waveform (as an estimate of signal amplitude) to the variance between individual sweeps at a single point (noise) (Elberling & Don, 1984, 1987a). The standard deviation ratio (SDR), on the other hand, estimates the signal-to-noise ratio from the standard deviation of the average waveform and the standard deviation of the (±) reference (Picton et al., 1983). The (±) reference is an estimate of the contribution of background noise to the average waveform. It is calculated by alternately adding and subtracting all of the recorded waveforms in a replication, then dividing by the number of trials (Schimmel, 1967). Values for each of these measures are reduced by excessive noise in the recording (Hyde, Sininger & Don, 1998), however, CCR has two additional weaknesses. Firstly, spurious peaks in the noise may replicate, and resulting in false positive identification of responses (Elberling & Don, 1987a). Secondly, CCR measures do not provide a RN estimate.

SDR, CCR and Fsp can be described as statistical measures insofar as they quantify and mathematically evaluate features of the ABR waveform. Moreover, criteria for these measures can be selected using either a statistical or an empirical approach (Elberling & Don, 1984; Hyde et al., 1998).. In the statistical approach, the probability estimate is based on a hypothetical distribution, and requires an estimate of the number

of independent data values. If this estimate is inaccurate, an inappropriate criterion may be selected (Hyde et al., 1998; Picton et al., 1983; Wong & Bickford, 1980).

SDR and CCR calculate variance over the time-domain of an average, such that the successive sample points are highly correlated. Moreover, at near-threshold intensities the response is smaller, and neighbouring sample points are likely to have similar amplitudes. Consequently, statistically-derived criteria of response significance would be relatively conservative (Picton et al., 1983), particularly in the case of threshold ABR. Fsp partially addresses the problem of point-to-point correlation by calculating the noise estimate across trials rather than over the response time window. However, 60-Hz line noise remains a potential source of correlation across sweeps, and may artificially inflate the estimate of signal amplitude or that of the noise measurement (Hyde et al., 1998). Because the variance of the average is calculated across the time domain, Fsp offers no advantage in terms of *df* in the signal estimate.

In light of the difficulties encountered in statistical analysis it seems appropriate to pursue empirically-determined cut-off values for response detection. An empirical criterion is one based on experimentally observed results. Signal detection theory (SDT) can be used to select an optimal criterion based on the sensitivity and specificity values which it yields. Sensitivity refers to the probability of a positive test outcome when the variable of interest (e.g., a response) is present. Specificity is the probability of a negative test outcome when that variable is absent (Hyde, Davidson, & Alberti, 1991). The actual presence or absence of the variable of interest is established using a gold-standard test, such as expert clinician judgement. A receiver operating

characteristic (ROC) curve can be generated and used to describe the overall test performance of the measure. Using ROC curves, examiners can also select an operating point based on their specific sensitivity and specificity requirements (Hyde et al., 1991).

In the case of ABR waveform classification, false negative errors (i.e., missing a response that is actually present) are less problematic than false positive errors (i.e., incorrectly stating that the waveform is response present, when there is no physiologic response to the stimulus). Although false negative errors could lead to over-estimation of a hearing loss, the decision-tree procedure for ABR testing provides a safeguard. When a clinician determines that there is no response to a stimulus, a later step is to acquire data for the same frequency at a higher intensity. The clinician then has latency and morphology information from the higher-intensity (i.e., suprathreshold) recording that can be used to re-evaluate the lower-intensity waveform. In contrast, false-positive errors are less likely to be detected within a test session than false-negative errors. This is because once a clinician decides that a response is present, testing for that frequency is complete. Importantly, the consequences of false-positive errors are potentially worse than those arising from false-negative errors. Hearing loss may be underestimated, or completely unidentified, leading to inadequate amplification until follow-up testing occurs as much as six months later. Furthermore, an unidentified hearing loss may impact speech-language, academic and social development (Carney & Møeller, 1998; Yoshinaga-Itano, Sedey, Coulter & Mehl, 1998, Yoshinaga-Itano, 2003). Based on the

above considerations, an optimal SDR criterion should aim to have near-perfect specificity (e.g., 95%), with reasonable sensitivity (e.g., 70% or higher).

Studies evaluating the test performance of quantitative feature detectors have shown promising results (e.g., Don, Elberling, & Waring, 1984; Elberling & Don, 1987a; Elberling & Don, 1987b; Norton, Gorga, Widen, Folsom, Sininger, Cone-Wesson, Vohr, Mascher, & Fletcher, 2000; Sanchez et al., 1995; Valdes-Sosa et al., 1987). Four important conclusions can be drawn from research on ABR analysis techniques. In the first place, the reliability of visual response identification depends on clinician experience (Valdes-Sosa et al., 1987). Furthermore, feature detectors in general have the ability to discriminate response presence and absence, with improved precision if the analysis window can be restricted to the region of Wave V (Sanchez et al., 1995). In addition, the use of empirically defined criteria for feature detectors may be appropriate in cases where statistical calculations underestimate the power of a test (Picton et al., 1983). Finally, high residual noise levels were found to increase detection errors (Valdes-Sosa et al., 1987).

Studies on the relationship between signal-to-noise ratio and response presence have concentrated almost exclusively on ABRs to click stimuli in normal-hearing subjects. Research is needed to determine whether, and to what extent, measures of signal-to-noise ratio are appropriate in the evaluation of brief-tone ABRs, in both normal-hearing and hearing-impaired populations.

The Intelligent Hearing Systems (IHS, Miami, Fl., U.S.A.) "Smart-EP" diagnostic ABR machine provides online SDR (and RN) calculations. Their measure is labelled SNR (for "signal-to-noise ratio") and its values have exactly half the magnitude of those

derived from SDR calculations. Pediatric audiology clinics throughout British Columbia have recently been furnished with the Smart-EP through the BC Early Hearing Program whose goal is early identification and intervention for hearing impaired infants (BCEHP, 2006). The objective of the present study is to define appropriate SNR criteria for the evaluation of ABR waveforms to 500-, 2000-, and 4000-Hz air-conducted stimuli from a sample of infants with normal or impaired hearing. If it is possible to establish a SNR criterion, response detection could conceivably become more accurate, more objective, and more time-efficient, resulting in more-complete diagnostic results. Improved diagnostic accuracy would help audiologists to avoid the pitfalls of either under- or over-amplification, and better outcomes could be expected.

Methods

Data collection

This study examined threshold-ABR data acquired at the Audiology Department of BC's Children's Hospital (BCCH) between January, 2005 and April, 2007. It consisted entirely of a chart review, with no additional testing of subjects required. This project was approved by the UBC Clinical Research Ethics Board, as well as the Children's and Women's Health Centre of British Columbia Research Review Committee.

Subjects

ABR results from 98 infants and young children were analysed. Subjects ranged in age from 20 days to 6.3 years (chronological age), with mean age at test of 1 year. Subjects were selected based on the availability of sufficient data, and without reference to presence, etiology, configuration, or degree of hearing loss. Data should therefore be representative of ABR participants at BCCH. Based on chart review, reasons for referral varied widely, from parental concern to syndromic differentiation. Data from the two subjects with suspected auditory neuropathy or auditory dyssynchrony were excluded in order to avoid confounds between the performance of ABR in auditory neuropathy, and the ability of SNR to identify an ABR response.

Audiology reports from the ABR test date indicate that 60% (n=59) of the 98 subjects had normal hearing in both ears, 8% (n=8) had conductive hearing loss in at least one ear, 8% (n=8) had a sensorineural impairment in at least one ear, and 2% (n=2) had a mixed loss in at least one ear. In the remaining 21% of cases (n=21), a

hearing impairment was diagnosed, but the type of hearing loss was not definitively stated. Some subjects were tested on multiple occasions and in some cases, results were available for one ear only. In all, 196 ears were tested, 143 of which had normal ABR thresholds. 35 of the 196 ears had a flat hearing loss (i.e., 500-, 2000- and 4000-Hz thresholds were all within 20 dB of one another), and 18 ears had a sloping hearing loss. ABR thresholds overall ranged from 10 dB nHL to no response at 110 dB nHL. In order to determine degree of hearing loss, estimated behavioural hearing level (EHL) corrections (minus 15 dB at 500 Hz, minus 5 dB at 2000 Hz, and no correction at 4000 Hz) (BCEHP, 2006; OIHP, 2005) were applied to ABR thresholds and a three-frequency average was calculated. 45 ears had average ABR thresholds at 25 dB EHL or better, 27 ears had thresholds between 26 and 45 dB EHL, five ears had thresholds between 46 and 65 dB EHL, seven ears had thresholds between 66 and 85 dB EHL, and 14 ears had ABR thresholds of 86 dB EHL or greater.

In order for the clinician to obtain adequate (and accurate) information during ABR testing, it was necessary for the child to remain asleep and quiet for an extended period of time (up to 3 hours). According to chart review, 13 of the subjects were tested under conscious sedation (using chloral hydrate), two were tested under general anesthesia, and the remaining 83 were tested in natural sleep.

Equipment/software

As per BCCH protocol, IHS Smart EP software was used for all ABR recordings.

Stimuli

Analyses focused on ABR data collected to 500-, 2000-, and 4000-Hz air-conducted stimuli. Stimulus parameters were set according to recommendations for brief-tone air-conducted stimuli (Stapells, 2000a). Specifically, the stimuli were "2-1-2" (cycles) linearly gated tones. The 500-Hz tone had 4-ms rise and fall times, with a 2-ms plateau. The 2000-Hz stimulus had 1-ms rise and fall times with a plateau of 0.5 ms. Finally, the 4000-Hz tone had rise and fall times of 0.5 ms with a plateau of 0.25 ms. All stimuli were presented via EARTONE 3-A insert earphones with 0 dB nHL equivalent to 22 dB peak-to-peak equivalent (ppe) SPL at 500 Hz, 20 dB ppe SPL at 2000 Hz, and 26 dB ppe SPL at 4000 Hz. Brief-tones were presented at a rate of approximately 39.1/s and alternating onset polarities were used.

All test-related decisions such as choice of frequencies/intensities, test sequence, and the number of sweeps acquired were made by the clinicians performing the test. Test protocol generally followed BCEHP recommendations (BCEHP, 2006).

Electroencephalogram (EEG) recordings

Each subject was fitted with four recording electrodes, with non-inverting at the vertex (or high forehead for young infants), inverting at the left and right mastoids, and ground at the forehead. Recording parameters were set according to the recommendations of Stapells (2000a), and interelectrode impedances were ≤ 5 kΩ.

Artifact rejection was set to exclude trials exceeding ±25 μV. The analysis time was 25.6 ms and a sampling rate of 20000 Hz was used.

Data analysis

Visual identification. Three expert judges rated previously obtained ABR results, such that each expert judge was asked to evaluate all of the waveforms based on visual observation. Two of the expert judges are clinicians who perform diagnostic brief-tone ABR daily, and who are mentor-trainers for BCEHP. The third judge is recognized worldwide as an expert in brief-tone ABR for clinical as well as research purposes. Judges were provided with paper copies of the waveforms, with each page containing three waveforms (a "set"). On the bottom of each page were two replications taken from a single subject and using the same stimulus. Each replication contained between 1786 and 7000 trials. Above them was a single waveform, representing the average of the two replications. The scale of waveform presentation was set to make small-amplitude responses easier to identify.

Waveform sets were organized in a binder according to stimulus frequency (500, 2000 or 4000 Hz). Within a frequency, the order of presentation of each set was randomized, such that consecutive pages did not contain waveforms that are related to each other (i.e., they were likely obtained from a different child, at a different intensity, and a different frequency). Moreover, the order of frequencies and the order of waveform sets at each frequency were randomized across judges.

Judges were asked to review each set of waveforms, and evaluate the presence/absence of a response. The judges rated the waveforms on a scale of 1-4, where 1 = no response, 2 = possible response, 3 = probable response and 4 = definite response (Nousak & Stapells, 2005; Stapells, Picton, Durieux-Smith, Edwards, & Moran, 1990). Each waveform set was then assigned an overall score, based on the average of the three judges' ratings. Waveforms with an average score of 2.5 or greater were deemed response present, and waveforms with average scores of less than 2.5 were deemed response absent.

Judges did not have knowledge of the intensity or frequency of the stimulus, or of the subject's hearing status, nor did they have access to values for objective measures such as RN, CCR or SNR. In addition, judges were blind to the opinions of the original clinicians.

Two indices of interrater reliability were calculated. The first index defined reliability in terms of differences in the number of categories between judges ratings (i.e., how often were all three judges ratings either the same or only one category apart). For all of the waveform ratings, judges were in agreement, within one category, in 98% of cases. In 49% of cases, all three judges gave the same rating to a waveform set. Reliability was also described as the proportion of cases in which all judges agreed whether a response was present or not. For those waveforms whose average score was 2.5 or greater, judges agreed 91% of the time that a response was present. For those with average scores below 2.5, there was 79% agreement. Overall, judges agreed on response presence or absence in 87% of cases.

SNR analysis. The IHS Smart EP system alternately allocates trials between two buffers,"A" and "B", such that each buffer comprises one half of the total number of trials. According to the IHS Smart EP algorithm, SNR is equal to the standard deviation of the average of the two buffers divided by the standard deviation of the difference between them. Standard deviations are calculated across the 10-ms SNR window. The SNR value was calculated for each set of waveforms, according to the equation:

$$SNR = \frac{SD_{(10\ ms)}\ \frac{(A+B)}{2}}{SD_{(10\ ms)}\ (A-B)}$$

A pilot study was performed in order to determine the most appropriate SNR window for each frequency, while avoiding stimulus artifact (see Appendix A for details). Based on results from the pilot study, SNR calculations were made using time windows of 10.5-20.5 ms, 6.5-16.5 ms, and 5-15 ms at 500, 2000 and 4000 Hz respectively.

Statistical analysis of RN and SNR. A three-way repeated-measures analysis of variance (ANOVA) was performed on SNR values. Factors were: Frequency (3 levels), Ear (2 levels), and Response Presence (2 levels). In cases where a subject had multiple recordings for a given category (e.g., two instances of response present recordings at 500 Hz in the left ear), one set of values was randomly selected. A three-way repeated-measures ANOVA was also carried out for RN values. Factors were the same as those described for SNR. Greenhouse-Geisser corrections for repeated measures (Greenhouse & Geisser, 1959) were applied to all analyses. ANOVAs were performed

using STATISTICA software (STATISTICA 6.1, 1984-2003) and results were considered significant at $p<.05$ for all tests.

ROC analysis. Non-parametric ROC curves were generated using the full range of SNR values in the data set[1]. SNR criteria were applied to each waveform, and ABRs classified as response present or absent accordingly. For each criterion, "hits" were defined as true responses (as per expert rating) that were correctly identified by the SNR measure; "false alarms" were the mistaken identification of responses where none existed. In order to determine the sensitivity and specificity of each SNR criterion, ROC curves were plotted for each test frequency (500 Hz, 2000 Hz, 4000 Hz) and for the three frequencies combined. On each curve, two operating points were identified: one for the SNR value which yielded 95% specificity and one for the SNR value which yielded 95% sensitivity.

In order to investigate the benefits of imposing noise limitations, additional ROC curves were plotted using a range of RN cutoffs (0.06 μV to 0.12 μV). Furthermore, ROC curves were plotted for each test frequency, given a RN cutoff of 0.08 μV.

For each ROC curve, area-under-the-curve and 95% confidence intervals were calculated. Area-under-the-curve represents the ability of a test to distinguish presence from absence in the variable of interest (in this case, an ABR response) (Hyde et al.,

---

[1]For presentation purposes, the ROC curves presented in the figures were generated using a range of SNR criteria from 0.3 to 2.00 with bin widths of 0.99. Thus, the bin labelled "0.3" included SNR values from 0.25 to 0.349. However, all test-performance, sensitivity, and specificity calculations as well as comparisons across tests were obtained from the original, non-parametric ROC curves.

1991). Area-under-the-curve values were compared statistically to determine the relative performance of the SNR measure for different conditions (e.g., for different RN exclusion criteria, and for 500 Hz versus 2000 Hz and 4000 Hz). StatsDirect software (StatsDirect Ltd., 1990-2007, http://www.statsdirect.com) was used for all ROC calculations.

Validation

Once criteria were chosen, results were validated on an additional sample of waveforms from chart review on 10 BCCH patients (not included in the 98 patients noted above). These 10 patients ranged in age from two months to 3.75 years, with a mean age at test of 11 months. Of the 20 ears tested, 14 had normal ABR thresholds, four had a flat loss, and two had sloping hearing loss. Four ears had average ABR thresholds between 26 and 45 dB EHL, one ear had an average threshold between 46 and 65 dB EHL, and one had an average threshold between 66 and 85 dB EHL.

The optimum criteria selected from the original sample of 98 patients was applied to the validation sample, and sensitivity and specificity values were calculated.

Results

Sample waveforms for 500-, 2000-, and 4000-Hz stimuli are presented in Figure 1. These data are drawn from 3 representative infants (one subject per frequency) with sensorineural hearing loss. An intensity series is displayed for each frequency, including response-present as well as response-absent waveforms. Also shown are the SNR and RN values for each waveform set (i.e., for the average of the replications), as well as the threshold determined by the judges blind to each subject's overall results. For those recordings judged response-present, SNR values were greater than 1.10, and SNR decreased with decreasing stimulus intensity (i.e., at 4000 Hz). Response-absent recordings had SNR values less than or equal to 0.86. RN values for all recordings in this sample were below 0.08 μV, and RN level was not affected by stimulus intensity.

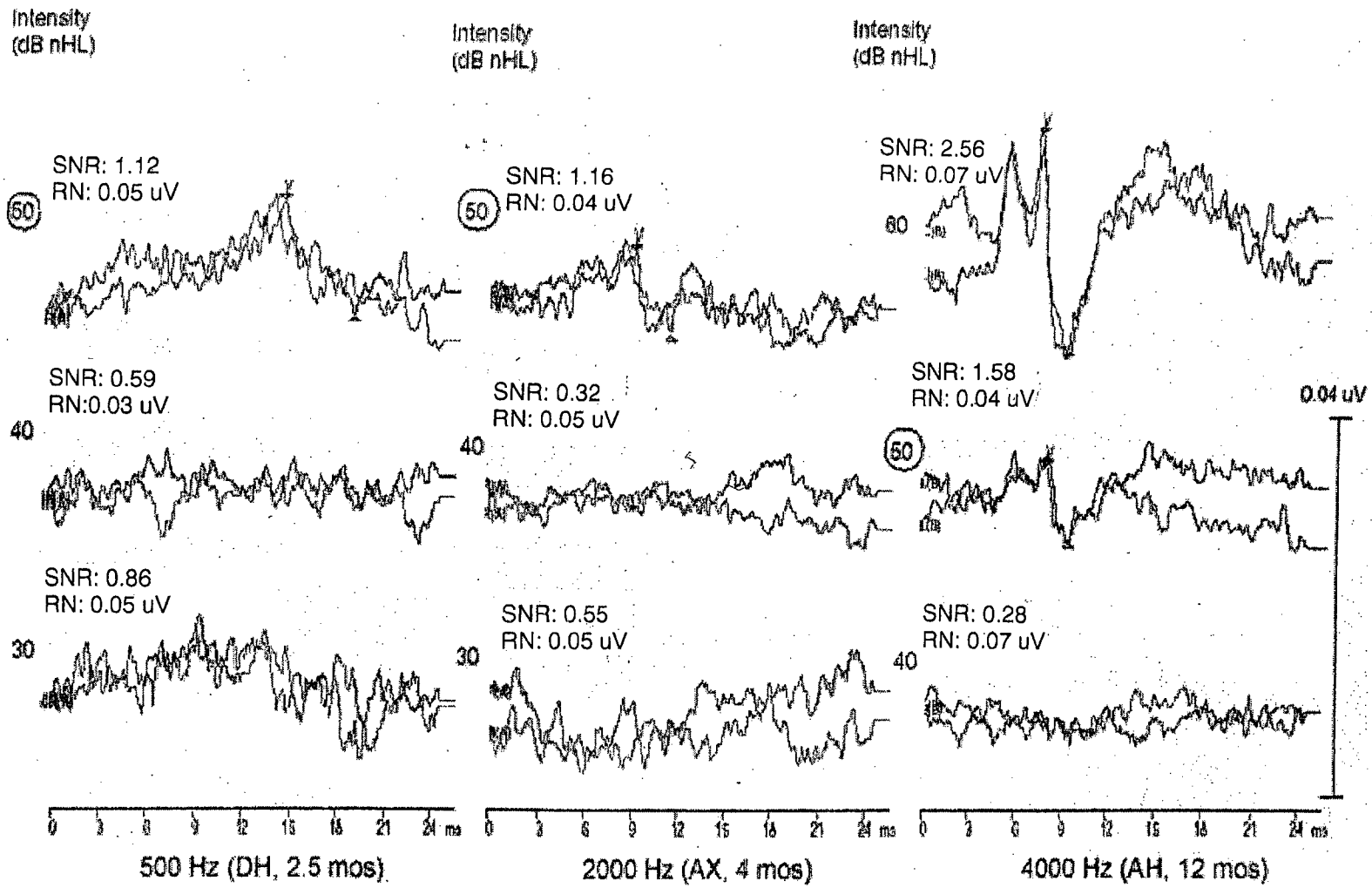

Figure 1. *Waveforms from three subjects (one per frequency) with sensorineural hearing loss. ABR thresholds, as determined by expert judges, are circled.*

RN and SNR values

According to expert ratings, 673 of the 937 waveforms were response present, and 264 were response absent. Table 1 provides the range of SNR values obtained across the 98 subjects in the original data set. Mean and median values are also provided. Table 2 shows the same data for RN values. For those waveforms which were judged response present, the mean SNR was 1.30, compared with a mean of 0.53 for response-absent waveforms and this difference was significant (see Table 3). Furthermore, 98% of response-absent waveforms had a SNR less than or equal to 1.0. SNR values decreased as stimulus frequency decreased, however, these differences were not significant. There was a significant effect of test ear, with the left ear showing slightly higher SNR values than the right ear. Mean SNR values for the left ear were 1.14, versus a mean SNR of 1.18 for recordings from the right ear. This difference is not meaningful for clinical purposes, and there was a great deal of overlap between the two distributions. There were no significant interactions between ear and response or between test ear and stimulus frequency.

Table 1.

SNR values by stimulus frequency

Frequency	Expert Rating	# Waveform Sets	Mean SNR	Median SNR	Standard Deviation	SNR Range
All Stimuli	RP	673	1.30	1.16	0.686	0.25-7.61
	NR	264	0.53	0.50	0.208	0.09-1.38
500 Hz	RP	214	1.13	1.04	0.531	0.28-3.03
	NR	89	0.52	0.50	0.182	0.18-0.92
2000 Hz	RP	242	1.33	1.18	0.719	0.27-7.61
	NR	114	0.52	0.49	0.210	0.09-1.27
4000 Hz	RP	217	1.43	1.28	0.750	0.25-4.14
	NR	61	0.58	0.56	0.230	0.24-1.38

Table 2.

RN values by stimulus frequency

Frequency	Expert Rating	# Waveform Sets	Mean RN (μV)	Median RN (μV)	Standard Deviation (μV)	RN Range (μV)
All Stimuli	RP	673	0.0702	0.0623	0.0328	0.0236-0.2960
	NR	264	0.0631	0.0541	0.0340	0.0227-0.3929
500 Hz	RP	214	0.0732	0.0661	0.0320	0.0246-0.2292
	NR	89	0.0636	0.0544	0.0336	0.0231-0.2564
2000 Hz	RP	242	0.0696	0.0621	0.0323	0.0236-0.2384
	NR	114	0.0675	0.0573	0.0394	0.0325-0.3929
4000 Hz	RP	217	0.0678	0.0601	0.0339	0.0248-0.2960
	NR	61	0.0545	0.0507	0.0193	0.0227-0.1228

Table 3.

Repeated measures ANOVA for SNR values

Effect	SS	DF	MS	F	P	G-G Epsilon
Intercept	193.7176	1	193.7176	283.5	0.0001	
Response	36.0748	1	36.0748	52.8	0.0001	
Error	53.9899	79	0.6834			
EAR	1.3135	1	1.3135	4.0	0.0477	1.000
EAR*Response	0.0099	1	0.0099	0.0	0.8620	1.000
Error	25.6465	79	0.3246			
FREQ	0.5460	2	0.2730	1.0	0.3643	0.8851
FREQ*Response	0.2707	2	0.1353	0.5	0.6052	0.8851
Error	42.4469	158	0.2687			
EAR*FREQ	0.0093	2	0.0046	0.0	0.9845	0.9882
EAR*FREQ*Response	0.3089	2	0.1545	0.5	0.5958	0.9882
Error	46.9717	158	0.2973			

The response-present waveforms in the sample had a mean RN of 0.0702 μV, and response-absent waveforms had a mean of 0.0631 μV. There was no significant difference between RN levels for response-present versus response-absent recordings at any test frequency (see Table 4). RN levels increased slightly as stimulus frequency decreased (i.e., recordings to 500-Hz stimuli tended to be noisier than recordings to 4000-Hz stimuli), but these differences between frequencies were not significant.

Table 4.

Repeated measures ANOVA for RN values

Effect	SS	DF	MS	F	P
Intercept	0.9459	1	0.9459	344.8	0.0001
Response	0.0078	1	0.0078	2.8	0.0960
Error	0.2167	79	0.0027		
EAR	0.0024	1	0.0024	2.5	0.1163
EAR*Response	0.0020	1	0.0020	2.1	0.1475
Error	0.0751	79	0.0009		
FREQ	0.0004	2	0.0002	0.3	0.7225
FREQ*Response	0.0002	2	0.0001	0.2	0.8372
Error	0.1083	158	0.0007		
EAR*FREQ	0.0000	2	0.0002	0.3	0.9661
EAR*FREQ*Response	0.0004	2	0.0002	0.3	0.7447
Error	0.1158	158	0.0007		

Test performance by frequency

Figure 2 shows ROC curves for the SNR data at 500, 2000, and 4000 Hz, and for all three frequencies pooled. In all cases, the areas-under-the-curve were greater than or equal to 0.89, indicating that SNR may be an appropriate measure for distinguishing response present waveforms from those with no response. Test performance was poorer for 500 Hz (*A* = 0.89) than for 2000 Hz (A=.93), and 4000 Hz (A=.91), but these differences were not significant (for 500 Hz versus 2000 Hz, p = 0.084; for 500 Hz versus 4000 Hz, p = 0.168).

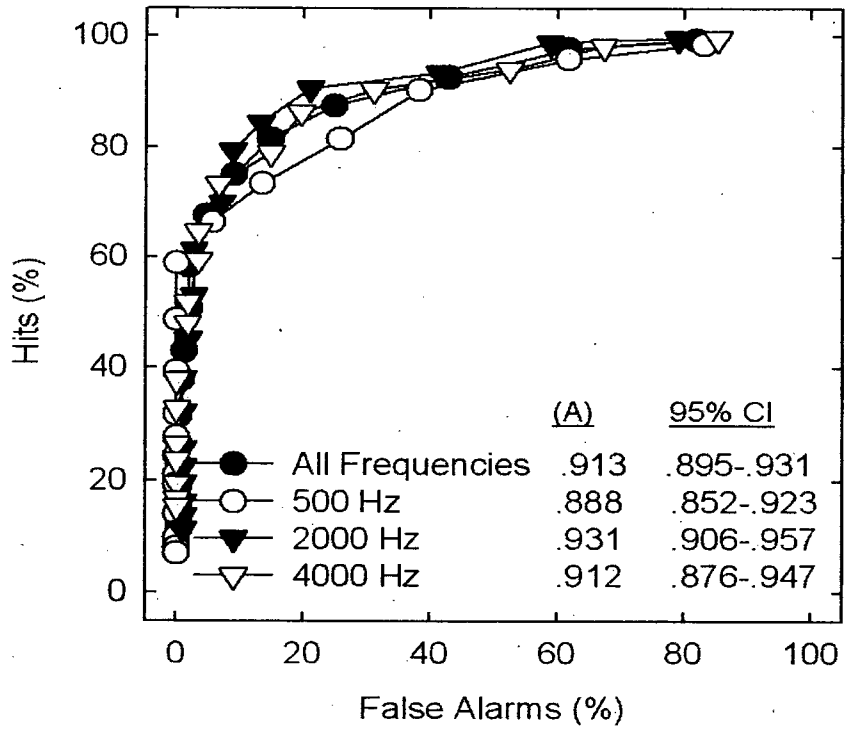

Figure 2. *ROC analysis of SNR test performance for 500 Hz, 2000 Hz, and 4000 Hz, as well as all three frequencies pooled. Area-under- the-curve values (A), and 95% confidence intervals are provided.*

For each ROC curve, two SNR values were calculated: (a) the SNR criterion corresponding to 95% sensitivity, and (b) the SNR criterion corresponding to 95% specificity. The range of criteria falling between these two SNR values represents the region of uncertainty where one cannot be certain whether a response is present or absent. Finally, the percent of total responses falling within the region of uncertainty was calculated. Table 5 lists the above data for each ROC curve. Table 6 outlines the relationship between sensitivity and specificity for each ROC curve. Results indicate that, for all frequencies combined, an SNR criterion of 0.90 yields 95% specificity, and 71% sensitivity. Thus, if the SNR value obtained for a waveform is greater than or equal to 0.90, the clinician can be reasonably certain that a response is in fact present. For SNR values less than 0.90, the clinician can say with 71% confidence that the waveform does not contain a response. Conversely, 95% sensitivity is achieved using a criterion of 0.51. Given a SNR value less than 0.51, a clinician can be reasonably confident that the waveform does not contain a response. This operating point yields a specificity of only 52%, however, and a SNR value greater than 0.51 is almost equally likely to be response present or response absent. 28% of the waveforms in this study had SNR values between 0.51 and 0.90, and therefore fell in the region of uncertainty.

Results suggest a difference in optimal SNR criteria for recordings to low-frequency versus higher-frequency stimuli. At 500 Hz, the operating point corresponding to 95% specificity was 0.86, versus 0.98 and 0.96 at 2000 and 4000 Hz respectively.

Table 5.

Evaluation of SNR Criteria.

Frequency	RN Cutoff (μV)	SNR Criterion		Proportion in Range of Uncertainty (%)
		95% Specificity	95% Sensitivity	
All Stimuli	None	0.90	0.51	28
	0.08	0.93	0.58	25
500 Hz	None	0.86	0.47	35
	0.08	0.86	0.52	28
2000 Hz	None	0.96	0.52	30
	0.08	0.98	0.66	21
4000 Hz	None	0.96	0.54	27
	0.08	0.96	0.67	17

Table 6.

The relationship between sensitivity and specificity by stimulus frequency and RN cutoff

Frequency	RN Cutoff (μV)	SENSITIVITY AT 95% SPECIFICITY (%)	SPECIFICITY AT 95% SENSITIVITY (%)
All Stimuli	None	72	52
	0.08	76	57
500 Hz	None	66	43
	0.08	74	49
2000 Hz	None	70	58
	0.08	76	75
4000 Hz	None	73	48
	0.08	79	70

Test performance and RN cutoffs

Three measures were used to evaluate the effect of RN cutoff on the diagnostic accuracy of SNR. Figure 3 demonstrates changes in test performance for all frequencies combined, over a range of RN exclusion criteria. Area-under-the-curve values varied from A=.93 when trials with RN >0.08 μV were excluded, to A=.91 when all waveforms were included. Overall, test performance improved as the RN cutoff neared 0.08 μV and only the 0.08 μV cutoff resulted in a significant improvement over no RN cutoff. When high-noise waveforms were excluded (e.g., RN >0.10 μV), fewer responses fell within the range of uncertainty (down from 28% to 25%). However, employing the strictest RN cutoff (0.06 μV) yielded slightly poorer results than the higher RN cutoffs. For all frequencies combined, sensitivity at the optimum criterion (i.e., 95% specificity) improved as the RN cutoff neared 0.08 μV. There was no systematic relationship between specificity at the 95% sensitivity criterion and RN cutoff level, although there was an improvement when very high-noise trials (RN >0.12 μV) were excluded.

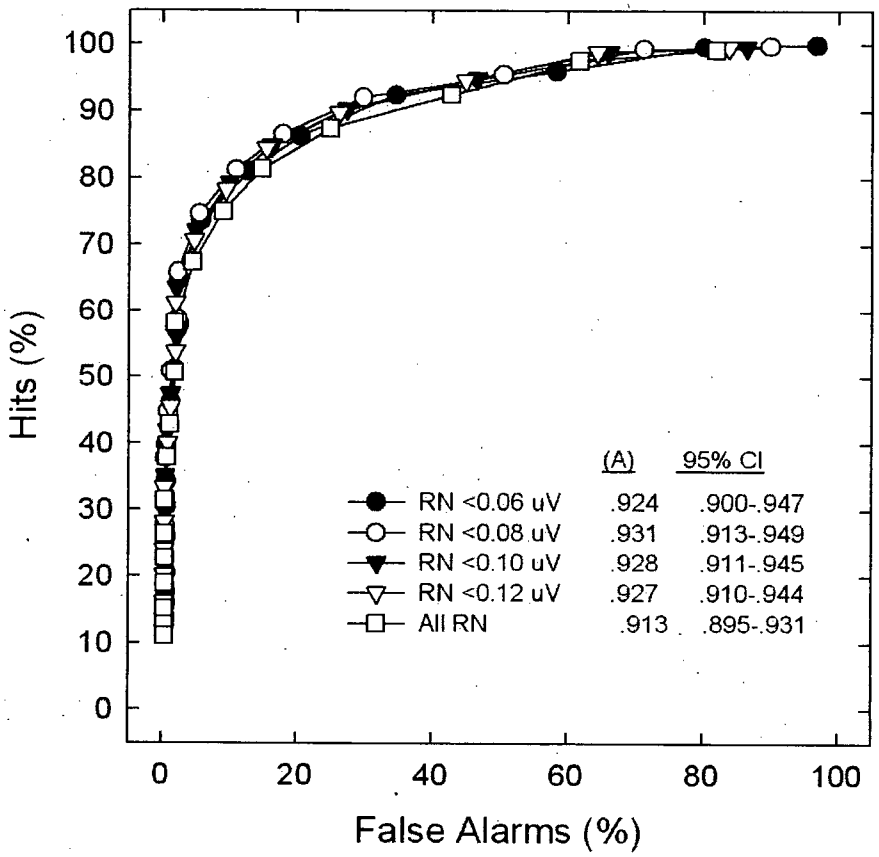

Figure 3. *Percentages of hits versus false alarms using SNR for recordings to 500-, 2000- and 4000-Hz stimuli pooled. ROC curves are plotted for a range of RN exclusion criteria. Test performance (A) and 95% confidence intervals are provided for each curve.*

Based on the above results, a RN cutoff of 0.08 μV was selected as the most appropriate. This cutoff was used to investigate the benefits of limiting noise levels at each of the test frequencies. Figure 4 displays comparisons of ROC curves with and without the RN cutoff, for each frequency. The data at each frequency were insufficient

for a statistical comparison of test performance between curves with or without a RN cutoff. However, visual inspection of the ROC curves suggest that test performance at each frequency improved with the exclusion of high-noise recordings (RN> 0.08 μV) yielding area-under-the-curve values of 0.94 at each of the three test frequencies. The largest improvement appears to be at 500 Hz.

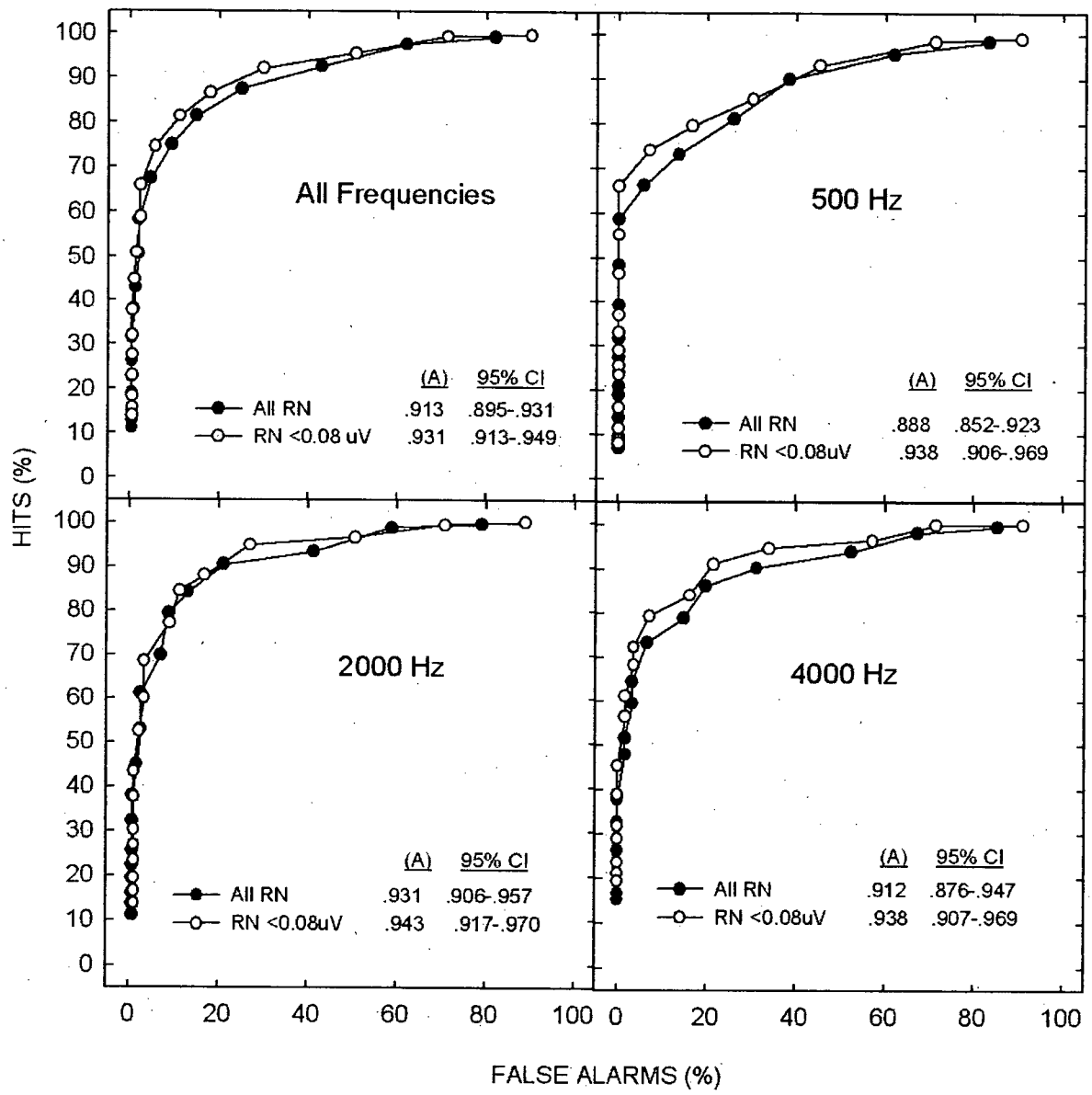

Figure 4. *ROC analyses of SNR for quiet recordings(RN<0.08* μV*) versus all recordings (no RN exclusion criterion) at 500 Hz, 2000 Hz and 4000 Hz, as well as all three test frequencies pooled. Areas-under-the-curve (A), and 95% confidence intervals are provided for each ROC curve.*

For low-noise waveforms (RN ≤0.08 μV), the SNR criterion corresponding to 95% specificity was 0.86 at 500 Hz, 0.98 at 2000 Hz, and 0.96 at 4000 Hz. Table 6 indicates that, at each frequency, sensitivity at the 95%-specificity criterion was higher for low-noise waveforms than when high-RN recordings were included in the sample. Moreover, the proportion of waveforms in the range of uncertainty decreased with the exclusion of high-noise waveforms. This result confirms the hypothesis that noisy recordings are more difficult to interpret than quiet recordings using SNR, particularly for 500 Hz.

Validation of SNR criteria

Table 7 depicts the sensitivity and specificity values obtained when the SNR criteria from Table 3 were applied to a validation sample of 10 subjects. For all data subsets, the criterion corresponding to 95% sensitivity in the original sample yielded at least 95% sensitivity in the validation sample. With the exception of 500 Hz, specificity values between the two samples corresponded well. For all stimuli (without a RN cutoff) the SNR criterion of 0.90 yielded specificity of 92% (as opposed to 95% in the original sample). At 2000 Hz and 4000 Hz, the pre-selected SNR criteria resulted in 100% specificity in the validation sample. At 500 Hz (with a RN cutoff of 0.08 μV), however, SNR criteria selected from the original sample yielded only 75% specificity in the validation sample. This indicates that the original criterion may have been too low. In

the validation sample of 500-Hz low-noise recordings, a criterion of 1.04 was required to achieve 95% specificity.

Table 7.

Validation of SNR criteria on an additional sample of 10 subjects (95 waveforms)

Frequency	RN Cutoff (μV)	Specificity at Original 95% Specificity Criterion (%)	Sensitivity at Original 95% Sensitivity Criterion (%)
All Stimuli	None	92	96
	0.08	91	98
500 Hz	None	59	95
	0.08	75	100
2000 Hz	None	100	96
	0.08	100	95
4000 Hz	None	100	100
	0.08	100	100

Finally, Table 8 shows sensitivity and specificity values for low-noise recordings from the original 98 subjects using SNR criteria between 0.96 and 0.98. Variations in sensitivity and specificity are small within this range (less than 4%), and no criterion performs best for all stimuli. Only the 0.98 criterion ensures a minimum of 95% specificity at all three frequencies (including, importantly, 500 Hz), and there is no specificity advantage to using a criterion of 0.97. As a result, SNR $\geq$0.98 appears to be the most appropriate criterion.

Table 8.

Sensitivity and specificity for low-noise recordings (RN≤0.08 μV) from the original 98 subjects, at a range of SNR criteria

SNR Criterion	500 Hz		2000 Hz		4000 Hz		All Stimuli	
	Spec. (%)	Sens. (%)	Spec. (%)	Sens. (%)	Spec. (%)	Sens. (%)	Spec. (%)	Sens. (%)
0.98	100	64	96	76	95	79	97	74
0.97	100	64	94	76	95	79	96	74
0.96	100	66	93	77	95	79	96	74

Discussion and Clinical Implications

From Tables 1 and 2, it is evident that a wide range of data was included in the present study. A large number of waveform sets was included at each stimulus frequency, and for each frequency at least 60 of the waveform sets were response absent. In addition, a wide range of SNRs were included in the sample. Current BCCH (and BCEHP) protocols emphasize the importance of minimizing noise levels, and RN values in the data set may be lower than typical levels at other clinics. Because the SNR measure has better performance for low-noise waveforms, results for those tests which included all RN levels may have been biased towards better performance because of relatively low RN levels. Importantly, there was no significant difference between RN values for response-present versus response-absent recordings. Employing RN cutoffs, therefore, did not alter the proportion of waveforms which contained a response. Overall, these results indicate that ROC analyses were based on a comprehensive data set.

Test performance

Test performance overall for the SNR measure was high, ranging from A=.89 to A=.94, depending on stimulus frequency and whether or not high-noise trials were excluded. As expected, test performance was poorer at 500 Hz than at 2000 Hz and 4000 Hz, although these differences were not significant. When high-noise waveforms were excluded, however, SNR was just as accurate at 500 Hz as for the other frequencies. It can be concluded, therefore, that responses to 500 Hz stimuli are more difficult to detect in higher background EEG noise (e.g., as a result of morphological

differences in responses to different frequencies). As long as noise levels are regulated, SNR performance is equal to A=.94 at each stimulus frequency. These results compare favourably with the other empirical studies (Arnold, 1985; Sanchez et al., 1995; Valdes-Sosa et al., 1987). These earlier studies based test performance calculations on pre- and post-stimulus intervals (as opposed to a continuum of sub- and super-threshold stimuli), and studied only ABRs to click stimuli. As noted by Valdes-Sosa et al., (1987), response detection is considerably more difficult at threshold, or near-threshold intensities. Despite the large proportion of ABRs to near-threshold or low-intensity stimuli in the sample, our study showed a high level of agreement between the SNR measure and expert ratings. Valdes-Sosa et al., (1987) found that experienced clinicians were even more accurate than statistical detection measures, whereas those with less experience performed significantly worse, even for suprathreshold recordings to click stimuli. Extrapolating from these findings, it is likely that the performance of the IHS SNR measure would be superior to that of new or inexperienced clinicians.

In the Norton et al., (2000) study, area-under-the-curve values are presented graphically for transient-evoked otoacoustic emissions, high- and low-level distortion product otoacoustic emissions (DPOAE), and ABRs to 30 dB nHL clicks (using Fsp). All measures were compared against a gold-standard of follow-up behavioural testing. At 1000 Hz *A* values ranged from approximately .70 for low-level (65/50) DPOAEs, to nearly .90 for the ABR measure. At 2000 Hz, click-ABR appeared to have the poorest performance with *A*=.89, whereas TEOAEs performed best at $A \approx .95$. At 4000 Hz, *A* values ranged from approximately .82 for high-level DPOAEs to .92 for TEOAEs. In

comparison, the SNR measure in the present study has a relatively strong test performance with A=.94 at 500, 2000, and 4000 Hz.

RN exclusion criteria

The data in this study demonstrates the importance of minimizing RN levels during ABR acquisition. The selection of an appropriate RN cutoff depends on the tradeoff between test-time (RN levels are reduced by increasing the number of trials in the average) and test performance. Strict RN criteria lead to more accurate results, but require valuable testing time. This study used three measures to evaluate the relative benefit of RN cutoffs ranging from 0.06 μV (the strictest) to no cutoff at all. In the first place, changes in test performance with RN cutoff were tabulated. For all frequencies combined, the highest test performance was achieved by limiting the RN level to 0.08 μV. Secondly, the relationship between sensitivity and specificity was compared for each RN cutoff. Findings suggest that, at the operating point corresponding to 95% specificity, sensitivity is highest for RN cutoffs of 0.08 μV and 0.10 μV. RN cutoffs also affected the proportion of SNR values which fell within the range of uncertainty. Waveforms in the range of uncertainty cannot be adequately interpreted by SNR criteria alone. Because it is desirable to minimize the number of responses which require additional (i.e., visual) interpretation, the proportion of SNR values in the range of uncertainty was compared across RN cutoffs. By excluding very high-noise waveforms (RN >0.12 μV), this proportion was reduced. The use of more-strict cutoffs did not substantially affect the proportion of uninterpretable SNR values, and the strictest cutoff

(RN > 0.06 μV), led to an increase in the proportion of uninterpretable SNR values relative to the other RN cutoffs. On each of the three measures (test performance, relationship between sensitivity and specificity, and effect on the range of uncertainty), the RN cutoff of 0.08 μV was either the most or, nearly the most, effective. Based on these findings, as well as test-time considerations, 0.08 μV was selected as the optimum RN cutoff for tone-evoked ABR.

In cases where a response can be clearly seen, and where there is a high SNR (e.g., greater than 1.0), it may not be necessary to continue acquiring until the recording is quiet. Because 98% of response absent waveforms had final SNR values less than or equal to 1.0, it is logical to assume that a false-positive error is unlikely given a clear response and SNR >1.0.

SNR criteria

Ideally, it would be possible to find a single SNR criterion which perfectly distinguished between response present and response absent waveforms. Errors in diagnostic ABR have serious consequences. At worst, false positive errors can cause a hearing loss to go unidentified and untreated. At the very least, false positive errors may lead to insufficient amplification if a hearing loss is underestimated. False negative errors, on the other hand, are less problematic because the error is likely to be corrected upon further testing at higher intensities. Unfortunately, no single SNR criterion has perfect sensitivity and specificity in relation to expert ratings in this study.

For all frequencies combined, the sensitivity corresponding to 95% specificity is 71%, and the specificity corresponding to 95% sensitivity is 52%. For the most part, these figures improve when the data are separated by frequency (requiring different SNR criteria for each frequency), and high-noise waveforms are excluded. At best, however, sensitivity plateaus at between 74% (at 500 Hz and 2000 Hz) and 79% (at 4000 Hz), in order to maintain a specificity of 95%.

A conservative approach to response detection with SNR would be to define two distinct criteria corresponding to 95% sensitivity and 95% specificity, for each data set. At 4000 Hz, for example, a SNR criterion of 0.67 yields 95% sensitivity (with a RN cutoff of 0.08 μV). Thus, a clinician could say with 95% certainty that a waveform with SNR less than 0.67 (and low RN) does not contain a response. However, the clinician would only have 48% confidence that a waveform with a SNR greater than or equal to 0.67 does contain a response. In order to address this concern, a second SNR criterion, yielding 95% specificity, could be defined. At 4000 Hz, this corresponds to a SNR of 0.96. The clinician could now be 95% certain that all waveforms with SNR values greater than or equal to 0.96 are response present. This leaves a range of SNR values, between 0.67 and 0.95, where the clinician must use additional information, such as the pattern over a range of intensities to determine whether or not a response is present. Given 95% confidence on either side of this range, only 16.8% of responses in the current study require additional (i.e., visual) interpretation.

Depending on the prevalence of hearing loss cases, as well as the cost-benefit priorities at a given test site, some degree of error in the response detection paradigm

may be acceptable. In these cases, clinicians could use the SNR criterion corresponding to 95% specificity (for each frequency), with the understanding that some responses will be missed. As discussed in the Introduction, these responses are likely to be identified later in the testing process as information at higher intensities is obtained.

Based on the original data sample, the SNR criteria corresponding to 95% specificity were 0.86, 0.98, and 0.96 for low-noise recordings at 500, 2000 and 4000 Hz respectively. When these criteria were applied to the validation sample, 100% specificity was achieved for both 2000 and 4000 Hz stimuli. At 500 Hz, however, the 0.86 criterion was too low, and resulted in a specificity of only 75%. As a result of these findings, higher criteria should be selected for recordings to 500 Hz stimuli. The data in Table 6 suggests that a criterion of 0.98 could be used to distinguish response-present from response-absent waveforms at 500, 2000 and 4000 Hz, with at least 95% specificity. Because SNR values for 1000-Hz data are not expected to differ from SNRs to the other frequencies, it can be inferred that the 0.98 criterion is appropriate for 1000 Hz as well.

For the sample ABR waveforms presented in Figure 1, a SNR criterion of 0.98 accurately distinguishes response-present from response-absent recordings at 500, 2000, and 4000 Hz. Thresholds determined by the three expert judges and by SNR were 50 dB nHL at all three frequencies.

Recommendations for further research

Currently, visual response interpretation by a clinician is the standard means of determining whether a response is present. Accordingly, this study used visual interpretation by expert judges as the gold standard for comparison. As a result, however, it is not possible to determine whether SNR would perform better or worse than visual techniques, or to what extent the use of SNR guidelines would improve diagnostic accuracy. What this study does tell us is the extent of agreement between SNR values and visual response detection by an expert. A high test performance value, therefore, indicates that the SNR measure was able to predict which waveforms would be considered response present by an expert judge, and distinguish those from the waveforms which would be considered response absent. The value of this type of analysis is that it allows comparison with other response detection measures. Moreover, it allows us to determine how to structure SNR guidelines in order to best approximate the current gold standard, and help relatively inexperienced clinicians to perform ABR testing with the highest possible accuracy.

Further research is needed to evaluate how SNR guidelines would impact diagnostic accuracy for new clinicians, both in terms of consistency, and agreement with expert clinicians. Furthermore, it would be useful to look at instances of disagreement between SNR and expert raters, in order to define situations in which the SNR value may not be reliable. Empirical research is needed to compare other statistical tools (such as Fsp) with SNR for brief-tone diagnostic ABR. Similar studies should also be

carried out with ABR to bone-conducted stimuli, as the optimal SNR criteria may differ from those selected for air-conducted stimuli.

Clinical implications and recommendations

Implications will be considered in terms of the three hypotheses posited in the introduction to this paper.

1. Do SNR values accurately differentiate present from absent responses?

The IHS SNR measure is a useful tool in distinguishing waveforms which contain a response from those which do not. As a result of the high test performance achieved, It could be used to assist and/or train inexperienced clinicians in interpreting waveforms, and is an additional source of information in the arsenal of the ABR expert.

2. Is SNR equally accurate for stimuli of different frequencies? Can the same operating points be used for all of the frequencies?

When high-noise recordings are excluded, SNR is equally accurate at 500 Hz, 2000 Hz, and 4000 Hz. *A criterion of 0.98 can be used at all three frequencies to detect ABRs with 95% specificity.* Sensitivity is higher for 4000 Hz (79%) and 2000 Hz (76%) than for 500 Hz (64%). A SNR criterion of 0.98 would likely be appropriate for 1000 Hz as well, although this was not tested.

3. Does test performance improve when high-noise recordings are excluded from the analysis?

Every measure of test performance and accuracy showed a benefit to maintaining low noise-levels. Based on an analysis of a range of RN levels, *a RN cutoff of 0.08 μV is*

recommended, particularly if the clinician wishes to base response interpretation on the SNR value alone. In cases *where a response is clearly present (using visual detection), and SNR is greater than 1.0 , the clinician may stop acquiring* despite a high RN level.

Bibliography

The ABaer and AOAE hearing screening system user's and service manual (2005). Mundelein, Illinois: Bio-logic Systems Corporation.

Alpsan, D., & Özdamar, O. (1992). Auditory brainstem evoked potential classification for threshold detection by neural networks. I. network design, similarities between human-expert and network classification, feasibility. *Automedica,* 15, 67-82.

American Academy of Audiology. (2003). *Pediatric amplification protocol*. Retrieved May 1, 2007, from http://www.audiology.org/NR/rdonlyres/53D26792-E321-41AF-850-CC253310F9DB/O/pedamp.pdf.

American Speech-Language-Hearing Association. (2004). *Guidelines for the audiologic assessment of children from birth to 5 years of age*. Retrieved May 1, 2007, from http://www.asha.org/NR/rdonlyres/OBB7C840-27D2-4DC6-861B-1709ADD78BAF/O/v2GLAudAssessChild.pdf.

Arnold, S. A. (1985). Objective versus visual detection of the auditory brain stem response. Ear and hearing, 6(3), 144-150.

BCEHP Diagnostic Advisory Group. (2006). BC early hearing program: Diagnostic audiology protocol. Retrieved May 10, 2007, from http://www.phsa.ca/AgenciesServices/Services/BCEarlyHearingPrgs/ForProfessionals/ProtocolsStandards/Dxprotocols_November_9_2006RJ[1].pdf.

Beattie, R. C. (1998). Normative wave V latency-intensity functions using the EARTONE 3A insert earphone and the radioear B-71 bone vibrator. *Scandinavian Audiology*, 27(2), 120-126.

Bendat, J. S., & Piersol, A. G. (1971). *Random data: Analysis and measurement procedures*. New York: Wiley.

Carney, A. E., & Moeller, M. P. (1998). Treatment efficacy: Hearing loss in children. Journal of speech, language, and hearing research : JSLHR, 41(1), S61-84.

Dobie, R. A. (1993). Objective response detection. *Ear and Hearing*, 14(1), 31-35.

Don, M., Allen, A. R., & Starr, A. (1977). Effect of click rate on the latency of auditory brain stem responses in humans. *The Annals of Otology, Rhinology, and Laryngology*, 86(2 pt. 1), 186-195.

Don, M., & Eggermont, J. J. (1978). Analysis of the click-evoked brainstem potentials in man unsing high-pass noise masking. *The Journal of the Acoustical Society of America*, 63(4), 1084-1092.

Don, M., & Elberling, C. (1996). Use of quantitative measures of auditory brain-stem response peak amplitude and residual background noise in the decision to stop averaging. *The Journal of the Acoustical Society of America*, 99(1), 491-499.

Don, M., Elberling, C., & Waring, M. (1984). Objective detection of averaged auditory brainstem responses. *Scandinavian Audiology*, 13(4), 219-228.

Durieux-Smith, A., Edwards, C. G., Picton, T. W., & McMurray, B. (1985). Auditory brainstem responses to clicks in neonates. *The Journal of Otolaryngology.Supplement*, 14, 12-18.

Durieux-Smith, A., Picton, T. W., Bernard, P., MacMurray, B., & Goodman, J. T. (1991). Prognostic validity of brainstem electric response audiometry in infants of a neonatal intensive care unit. *Audiology : Official organ of the International Society of Audiology,* 30(5), 249-265.

Durrant, J. D., & Ferraro, J. A. (1999). Short-latency evoked potentials: Electrocochleography & auditory brainstem response. In F. E. Musiek, & W. F. Rintelmann (Eds.), *Contemporary perspectives in hearing assessment* (pp. 197-240). Boston: Allyn & Bacon.

Elberling, C., & Don, M. (1984). Quality estimation of averaged auditory brainstem responses. *Scandinavian Audiology*, 13(3), 187-197.

Elberling, C., & Don, M. (1987a). Detection functions for the human auditory brainstem response. *Scandinavian Audiology*, 16(2), 89-92.

Elberling, C., & Don, M. (1987b). Threshold characteristics of the human auditory brain stem response. *The Journal of the Acoustical Society of America*, 81(1), 115-121.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24, 95-112.*

Hecox, K., & Galambos, R. (1974). Brain stem auditory evoked responses in human infants and adults. *Archives of Otolaryngology* (Chicago, Ill.: 1960), 99(1), 30-33.

Hyde, M., Sininger, Y. S., & Don, M. (1998). Objective detection and analysis of auditory brainstem response: An historical perspective. *Seminars in Hearing*, 19(1), 97-113.

Hyde, M. L., Davidson, M. J., & Alberti, P. W. (1991). Auditory test strategy. In J. T. Jacobson, & J. L. Northern (Eds.), *Diagnostic Audiology* (pp. 295-322). Austin, Texas: Pro-ed.

Hyde, M. L., Riko, K., & Malizia, K. (1990). Audiometric accuracy of the click ABR in infants at risk for hearing loss. *Journal of the American Academy of Audiology*, 1(2), 59-66.

Jewett, D. L., Romano, M. N., & Williston, J. S. (1970). Human auditory evoked potentials: Possible brain stem components detected on the scalp. Science (New York, N.Y.), 167(924), 1517-1518.

Jewett, D. L. (1994). A janus-eyed look at the history of the auditory brainstem response as I know it. *Electromyography and Clinical Neurophysiology*, 34(1), 41-48.

Joint Committee on Infant Hearing. (2000). Year 2000 position statement: Principles and guidelines for early hearing detection and intervention programs. [Electronic version]. Pediatrics, 106(4), 798-817. Retrieved May 1, 2007, from http://www.jcih.org/jcih2000.pdf

Møeller, A. R., Jho, H. D., Yokota, M., & Jannetta, P. J. (1995). Contribution from crossed and uncrossed brainstem structures to the brainstem auditory evoked potentials: A study in humans. *Laryngoscope*, 105(6), 596-605.

Norton, S. J., Gorga, M. P., Widen, J. E., Folsom, R. C., Sininger, Y., Cone-Wesson, B., et al. (2000). Identification of neonatal hearing impairment: Evaluation of transient evoked otoacoustic emission, distortion product otoacoustic emission, and auditory brain stem response test performance. *Ear and Hearing*, 21(5), 508-528.

Nousak, J. K., & Stapells, D. R. (2005). Auditory brainstem and middle latency responses to 1 kHz tones in noise-masked normally-hearing and sensorineurally hearing-impaired adults. *International Journal of Audiology*., 44(6), 331-344.

Oates, P., & Stapells, D. R. (1997a). Frequency specificity of the human auditory brainstem and middle latency responses to brief tones. I. high-pass noise masking. *The Journal of the Acoustical Society of America*, 102(6), 3597-3608.

Oates, P., & Stapells, D. R. (1997b). Frequency specificity of the human auditory brainstem and middle latency responses to brief tones. II. derived response analyses. *The Journal of the Acoustical Society of America,* 102(6), 3609-3619.

Ontario Infant Hearing Program. (2005). Audiologic assessment protocol and support documentation. Retrieved 05/10, 2007, from www.mtsinai.on.ca/IHP/english/documents/IHPAssessmentProtocolRevisedfinaldraftAug2405.pdf.

Österhammel, P. A., Davis, H., Wier, C. C., & Hirsh, S. K. (1973). Adult auditory evoked vertex potentials in sleep. *Audiology : Official Organ of the International Society of Audiology*, 12(2), 116-128.

Picton, T. W. (1990). Auditory evoked potentials. In D. D. Daly, & T. A. Pedley (Eds.), *Current practice of clinical electroencephalography* (pp. 625-678). New York: Raven Press Ltd.

Picton, T. W., Hink, R. F., Perez-Abalo, M., Linden, R. D., & Wiens, A. S. (1984). Evoked potentials: How now? *Journal of Electrophysiological Technology*, 10, 177-221.

Picton, T. W., Durieux-Smith, A., & Moran, L. M. (1994). Recording auditory brainstem responses from infants. *International Journal of Pediatric Otorhinolaryngology*, 28(2-3), 93-110.

Picton, T. W., Linden, R. D., Hamel, G., & Maru, J. T. (1983). Aspects of averaging. *Seminars in Hearing*, 4, 327-341.

Picton, T. W., Stapells, D. R., & Campbell, K. B. (1981). Auditory evoked potentials from the human cochlea and brainstem. *The Journal of Otolaryngology*. Supplement, 9, 1-41.

Sanchez, R., Riquenes, A., & Perez-Abalo, M. (1995). Automatic detection of auditory brainstem responses using feature vectors. *International Journal of Bio-Medical Computing,* 39(3), 287-297.

Schimmel, H. (1967). The (+) reference: Accuracy of estimated mean components in average response studies. *Science*, 157(784), 92-94.

Schimmel, H., Rapin, I., & Cohen, M. M. (1974). Improving evoked response audiometry with special reference to the use of machine scoring. *Audiology : official organ of the International Society of Audiology*, 13(1), 33-65.

Shepherd, D. C., & McCarren, K. (1972). An averaged electroencephalic audiometric sensitivity (AEA-S) procedure. *The Journal of Speech and Hearing Disorders*, 37(4), 503-522.

Sininger, Y. S. (1993). Auditory brain stem response for objective measures of hearing. *Ear and Hearing*, 14(1), 23-30.

Sininger, Y. S., & Masuda, A. (1990). Effect of click polarity on ABR threshold. *Ear and Hearing*, 11(3), 206-209.

Stapells, D. R. (2000a). Frequency-specific evoked potential audiometry in infants. In R. C. Seewald (Ed.), *A Sound Foundation Through Early Amplification* (pp. 13-31). Basel: Phonak AG.

Stapells, D. R. (2000b). Threshold estimation by the tone-evoked auditory brainstem response: A literature meta-analysis. *Journal of Speech-Language Pathology and Audiology*, 24(2), 74-81.

Stapells, D. R., Picton, T. W., Durieux-Smith, A., Edwards, C. G., & Moran, L. M. (1990). Thresholds for short-latency auditory-evoked potentials to tones in notched noise in normal-hearing and hearing-impaired subjects. *Audiology*, 29(5), 262-274.

Takagi, N., Suzuki, T., & Kobayashi, K. (1985). Effect of tone-burst frequency on fast and slow components of auditory brain-stem response. *Scandinavian Audiology*, 14(2), 75-79.

Valdes-Sosa, M. J., Bobes, M. A., Perez-Abalo, M. C., Perera, M., Carballo, J. A., & Valdes-Sosa, P. (1987). Comparison of auditory-evoked potential detection methods using signal detection theory. *Audiology, Journal of Auditory Communication. Audiologie, Journal de la Communication Auditive*, 26(3), 166-178.

Weinberg, H., & Cooper, R. (1972). The recognition index: A pattern recognition technique for noisy signals. *Electroencephalography and Clinical Neurophysiology*, 33(6), 608-613.

Wicke, J. D., Goff, W. R., Wallace, J. D., & Allison, T. (1978). On-line statistical detection of average evoked potentials: Application to evoked response audiometry (ERA). *Electroencephalography and Clinical Neurophysiology,* 44(3), 328-343.

Wong, P. K., & Bickford, R. G. (1980). Brain stem auditory evoked potentials: The use of noise estimate. *Electroencephalography and Clinical Neurophysiology*, 50(1-2), 25-34.

Wu, C., & Stapells, D. R. (2001). Objective detection (fsp) of infants' auditory brainstem response (ABR) to clicks and brief tones. Retrieved April 10, 2007, from http://www.ausp.memphis.edu/ierasg/ierasg2001/Wu.txt.

Yoshinaga-Itano, C. (2003). Early intervention after universal neonatal hearing screening: Impact on outcomes. *Mental Retardation and Developmental Disabilities Research Reviews*, 9(4), 252-266.

Yoshinaga-Itano, C., Sedey, A. L., Coulter, D. K., & Mehl, A. L. (1998). Language of early- and later-identified children with hearing loss. *Pediatrics,* 102(5), 1161-1171.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clinical Chemistry,* 39(4), 561-577.

Appendix A: Pilot Study

Pilot Study: SNR Windows

The purpose of the pilot study was to define optimal SNR windows for 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz brief-tone stimuli. As outlined in the Introduction, signal-to-noise-ratio algorithms are significantly better at identifying a response when the calculation is restricted to a narrow time window around the expected response latency. Moreover, the latency at which a response is most likely to occur varies with the nominal frequency of the stimulus (e.g., wave V is typically earlier to a 4000 Hz stimulus than to a 500 Hz stimulus), and with the level relative to threshold (i.e., responses occur later to stimuli at threshold than at higher sensation levels). The IHS Smart EP system has a default SNR window of 4.00ms to 9.00ms (based on click-ABR); however clinicians are able to set this parameter to other times.

The full ABR window is typically set to 25.6 ms when recording to brief-tone stimuli. A 10-ms-wide SNR window would allow the clinician to hone in on the expected response latency, while maintaining the ability to detect responses which occur slightly earlier or later than the norm.

Data were taken from 30 of the subjects included in the larger study. These subjects varied in terms of the presence, type and configuration of hearing loss. Furthermore, a wide variety of waveform sets were selected to represent a range of intensities (20 - 100 dB nHL), as well as levels relative to threshold (i.e., many of the waveforms were at threshold, while others were well above threshold).

For each stimulus frequency (500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz) 50 sets of waveforms were printed. For each set of waveforms, the two original replications were overlapped and placed below the average waveform. In cases where there were more

than two replications, an average of all replications was taken and then split into two buffers (rarefaction and condensation). The rarefaction and condensation waveforms were overlapped and placed below the average waveform.

An expert rater was asked to identify the 10 ms window which optimally reflected the response for each waveform set. Consideration was also given to the exclusion of stimulus artifact. Stimulus artifact may artificially inflate SNR values if included in the SNR window. For this task, the rater was provided with a transparent ruler with a clearly labelled 10 ms window (to scale), and markings were made directly on the waveform. In some cases, several windows would have been equally appropriate (e.g., 6 ms-16ms, and 7ms - 17 ms windows both contained the entire response and minimal noise). All possible windows were included in the analysis. The mean SNR-window start time was tabulated for each stimulus frequency. Outliers (responses that were substantially earlier or later than the mean) were then re-evaluated. In order to determine wether their SNRs would be significantly affected by changing the window, SNRs were calculated using a variety of windows. For instance, SNR was calculated for the outliers at 500 Hz using windows of 7-17 ms, 9-19 ms, 10-20 ms, and 11-21 ms. The results of the pilot study are outlined in Table 9.

Table A.1.

Optimal time windows for the calculation of SNR

Stimulus Frequency (AC)	**SNR Window**
500 Hz	10.5 - 20.5 ms
2000 Hz	6.5 - 16.5 ms
4000 Hz	5 - 15 ms

Appendix B: CREB Ethics approval

Appendix C: BCCH Ethics approval