Published in final edited form as:

Ear Hear. 2012 ; 33(2): 267–278. doi:10.1097/AUD.0b013e318230bba0.

Auditory Steady State Responses in Normal-Hearing and Hearing-Impaired Adults: An Analysis of Between-Session Amplitude and Latency Repeatability, Test Time, and F Ratio Detection Paradigms

Timothy Wilding, Bsc,

timothy.wilding@postgrad.manchester.ac.uk Audiology and Deafness Research Group Ellen Wilkinson Building University of Manchester Oxford Road, Manchester M13 9PL Tel: +44 161 2758286

Colette McKay, Prof.,

colette.mckay@manchester.ac.uk Audiology and Deafness Research Group Ellen Wilkinson Building University of Manchester Oxford Road, Manchester M13 9PL Tel: +44 161 2761671

Richard Baker, PhD, and

richard.baker@manchester.ac.uk Audiology and Deafness Research Group Ellen Wilkinson Building University of Manchester Oxford Road, Manchester M13 9PL Tel: +44 161 2753388

Karolina Kluk, PhD

karolina.kluk@manchester.ac.uk Audiology and Deafness Research Group Ellen Wilkinson Building University of Manchester Oxford Road, Manchester M13 9PL Tel: +44 161 2753371

The repeatability of auditory steady-state response (ASSR) amplitudes has not been well examined in hearing-impaired subjects. In this study we determine the between session repeatability coefficients of ASSR amplitudes and latencies in normally hearing and hearing-impaired subjects.

INTRODUCTION

The between test session repeatability of auditory steady-state response (ASSR) amplitudes has received little attention, especially in hearing-impaired subjects. This could be an important factor in ASSR threshold estimation as the detection of a response relies on the response amplitude being sufficiently greater than the residual electroencephalographic (EEG) random noise in the recording. The repeatability of ASSR thresholds depends on the repeatability of response amplitudes and EEG-noise levels (Picton et al. 2003). In this study we aimed to examine the repeatability of ASSR amplitude and latency in both normally hearing and hearing-impaired subjects between two separate recording sessions. We also examined the response detection rate against test time and response detection parameters.

The ASSR is a form of envelope-following evoked potential. The response is recorded by measuring the electrical potential difference across two electrodes usually positioned at the vertex (or high forehead) and ipsilateral mastoid (or nape of the neck). The response is separated from the background EEG by averaging, and statistical analysis techniques are used to determine the presence or absence of a response. The ASSR can be used as a form of objective audiometry by determining the threshold level at which the response is detectable (Galambos et al. 1981). The ASSR is a continuous response generated from amplitude modulated (AM), frequency modulated (FM) or a simultaneous mix of both frequency and amplitude modulated tones (MM) (Picton et al. 2003). The EEG is time domain averaged and then converted to its spectral components using a Fast-Fourier-Transform (FFT)

technique. The FFT produces amplitude and phase information in discrete frequency bins. The amplitude and phase of the response frequency bin (the frequency bin corresponding to the modulation frequency of the stimulus) is then compared to the random EEG noise in the adjacent frequency bins. The response is deemed to be present if the response FFT bin is significantly different from the noise bins. There are several statistical tests used to determine the presence of the response. These statistical tests consider either response and noise amplitude (F-test), response phase (phase coherence), or both amplitude and phase (Hotelling's T^2 and magnitude squared coherence) (John et al. 2008). Valdes et al. (1997) compared four possible response detection statistical methods (coherence synchrony measures, circular T^2 , F-test and Hotellings T^2) and found that all the tests detected the presence of responses with equal efficiency. Dobie et al. (1996) compared the F-test against segment analysis using magnitude-squared coherence and found identical performance with both methods.

The F-test is utilized by the multiple auditory steady-state evoked response (MASTER, Rotman Research Institute, Toronto-Canada) system. In the F-test the frequency bin corresponding to the stimulus modulation frequency is examined and compared with the background EEG noise in the adjacent 120 FFT bins (60 above and 60 below) using the F-ratio statistical test. The statistical test produces a significance p-value, which is tested against the chosen significance level. In ASSR-threshold measurements the threshold is defined as the lowest signal level at which the ASSR amplitude is found to be significantly above the EEG-noise floor using the chosen F-test criterion. There are different possible detection protocols with differing F-test significance percentage levels, using fixed, or variable test time. In statistical testing, type I (falsely finding a difference) and type II errors (not finding a difference which is present) can occur. The chance of type I and type II errors depend upon the selected p-value. In ASSR detection using an F-test detection criteria of $p < 0.05$, incorrectly indentifying noise as response (i.e. false response detection) is expected to occur in 5% of recordings by chance. However, this false detection rate is increased when using variable test time protocols that stop testing once the required significance criterion is met. The increased error rate occurs since variable test time protocols repeatedly test the EEG data following each additional sequential segment of recording time (sweeps). The repeated testing increases the chance of a false rejection of the null-hypothesis (i.e type I error, false response detection). John and Purcell (2008) explained the principles of the possibility of the statistical errors that could occur using multiple testing. The chances of false response detection increase from 5% to 54% in fifteen multiple tests (i.e. a 15 sweep ASSR recording). It is possible to apply a Bonferroni adjustment to the F-test ratio p-level to reduce the chance of false response detection. However, applying an adjusted (lower) p-level can be problematic as it can prevent true response detection and the adjusted p-levels are too conservative as the consecutive cumulative averaged EEG samples are highly correlated (John and Purcell 2008; Sturzebecher et al. 2005). It is possible to correct for the correlation using a modified p-level adjustment calculation. At the 5% level, for 15 sweeps, the adjusted p-level assuming cumulative EEG sample correlation $r=0.75$ is $p=0.03$.

Luts et al. (2008) further examined the effect of different ASSR detection paradigms on the test time required for threshold estimation and false response detection rates. They determined that the response detection time could be reduced by using variable test time protocols. However, there was a tradeoff between the reduction of test time and an increase in detection rate of false response. The most extreme example of which was the variable test time with 5% level on F-test protocol (i.e. stop recording when 5% level reached), which had the shortest test times but the highest detection rate of false response (30% of the recordings). The false detections found using the 5% variable test time protocol occurred in two consecutive test runs (at differing signal levels), which could have led to false threshold in an unacceptably high number of cases (24%). Their work highlighted the importance of

considering both test time and error rates in choice of ASSR detection protocols. D'Haenens et al. (2010) reanalyzed their earlier (2008) data and performed additional recordings in hearing-impaired subjects. They found that the false response detection rates were at worst 41.2% for the variable test time protocol, i.e. stop recording when 5% significance reached. Both D'Haenens et al. (2010) and Luts et al. (2008), found that the error rates could be reduced by using detection protocols that required the response to remain significant for more than one consecutive sweep. At present commercially available ASSR systems do not offer the possibility of reducing error rates by using corrected p-levels, nor by using additional significance criteria such as those used by D'Haenens et al. (2010) and Luts et al. (2008).

The response amplitude is dependent on the stimulus modulation and carrier frequencies and the subject's state of wakefulness (Cohen et al. 1991). The source of the response depends on the modulation rate of the stimuli used to evoke the ASSR. The 40-Hz modulation-frequency range response, which is generated predominantly by the auditory cortex, gives a higher recorded response amplitude and is more highly affected by attention than the 80-Hz range response, which is generated by sub-cortical structures (Cohen et al. 1991; Picton 2010).

In ASSR threshold measurements between-recording-session differences in response or in noise can affect the repeatability of the measured threshold. Champlin (1992) examined the stability of the 40-Hz range responses and found that thresholds determined using statistical response detection methods that relied on response phase had a higher level of correlation in repeated recordings than those that relied on response amplitude, and concluded that response phase is more stable than response amplitude. Kaf et al (2006) determined correlation coefficients of between-session repeated ASSR-threshold measurements in the 80-Hz range from normally hearing female subjects and found moderately strong correlations for 1-, 2-, and 4-kHz carrier frequencies and poorer correlation at 500 Hz. However, the use of correlation coefficients to determine repeatability can be misleading as correlation measures the strength of relation between two variables not their agreement (Bland et al. 1986). D'Haenens et al. (2008) determined the repeatability of ASSR thresholds to be within ± 10 dB between sessions for 2-kHz stimuli in the 80-Hz modulation range. They calculated the standard error of measurement (SEM) of the recorded response amplitudes. The SEMs were estimates of the standard deviations of repeated measurements calculated from the standard deviations of each individual's measurements. The repeatability was reported as ± 2 SEM which were an estimate of the 95% confidence interval limits of the differences. The repeatability was found to be 15 nV (29% of the mean response amplitude) for a 2-kHz stimulus carrier frequency at 50 dB HL, which was the highest stimulus level they tested. The mean amplitude and absolute values of repeatability (± 2 SEM) decreased with decreasing stimulus level. However, as stimulus level decreased the ratio of the repeatability (± 2 SEM) to the mean amplitude increased and at levels near to mean threshold for the 2-kHz stimuli the repeatability was 80% of the mean response amplitude. The increase in repeatability (± 2 SEM) as a percentage of mean response amplitude is expected as the response amplitudes were not EEG-noise corrected. As the response amplitudes decreased the residual EEG-noise in the response frequency bin would become more dominant, leading to both amplitude and variability being determined largely by the EEG-noise level.

The study presented here adds to the current body of research in this area by determining the repeatability of response amplitudes in hearing-impaired as well as normally hearing subjects. Response amplitudes were chosen for analysis, rather than phase, as response amplitudes (F-test) are the recording analysis method of choice for the MASTER ASSR system. We also analyzed the test time and false response detection rates of our recorded

data using similar methods to Luts et al. (2008) and D'Haenens et al. (2010) with the aim of adding to the discussion of best practice ASSR detection protocols.

METHODS

Subjects

Ten male and ten female normally hearing adults, aged 18 to 49 years, were recruited from staff and students at the University of Manchester, United Kingdom. Nine male and two female hearing-impaired adult subjects, aged 60 to 77 years, were recruited from the hearing aid clinic of the Manchester Royal Infirmary Audiology Department. The hearing-impaired subjects were older than the normally hearing subjects as it was not possible to recruit age-matched hearing-impaired subjects. The study received ethical approval from the University of Manchester, School of Psychological Sciences Ethics Board and the North Manchester Research Ethics Committee.

Normally hearing subjects (N1-N20) had air conduction thresholds ≤ 20 dB HL at standard audiometric frequencies (250 Hz to 8 kHz) and had no history of otological or neurological pathology. The subjects received financial compensation for their time. The normally hearing subjects were the same subjects as reported in Wilding et al. (2011).

The hearing-impaired subjects were aided in both ears, had a symmetrical sensory-neural hearing loss, normal tympanometry, and no history of middle ear disease. The subjects were recruited in two groups to represent extremities of hearing impairments, i.e. "mild hearing loss" with thresholds between 20 and 40 dB HL at 2 kHz (M1-M5); and "severe hearing loss" with thresholds between 70 and 90 dB HL at 2 kHz (S1-S5). One severely hearing-impaired subject withdrew from the study after one session and was excluded from analysis and thus was not assigned a subject number.

Stimuli

The ASSR stimuli were presented through ER3A (Etymotic Research Incorporated, Grove Village-Illinois) insert earphones. The stimuli were digitally generated using a MATLAB (The Mathworks Incorporated, Natick-Massachusetts) program with a sample rate of 32 kHz. The ASSR test stimulus was an exponentially amplitude modulated (AMEXP) 2-kHz carrier tone. The amplitude modulation depth was set to 100%. AMEXP stimuli have a wider signal bandwidth than AM alone but elicit higher ASSR amplitudes at low sensation levels compared to AM, maximizing the ratio of response amplitude to EEG noise (John et al. 2002). The ASSR modulation frequency was set to 95 Hz, as Cohen et al. (1991) found that the response amplitude peaks near to this modulation rate in adults.

The ASSR stimuli were imported into the MASTER system and presented to the insert earphones via the external input channel of a GSI61 (Grason-Stadler, Eden Prairie-Minnesota) audiometer. Calibration was performed using a GRAS type 26AC .25" (G.R.A.S. Sound and Vibration A/S, Holte-Denmark) microphone connected to Agilent 35670A Dynamic Signal Analyzer (Agilent Technologies UK Ltd, Wokingham, United Kingdom) in octave analysis mode via a GRAS IEC711 coupler (G.R.A.S. Sound and Vibration A/S, Holte-Denmark). In this experiment all stimuli were calibrated in dB SPL.

Procedures

Subjects attended two test sessions, which were conducted on separate days. The subjects' skin was prepared using Nuprep (Weaver and Company, Aurora-Colorado) abrasive cleaner and single-use electrodes were attached to the vertex, nape of the neck just below the hair line and right collarbone (ground) using TEN20 (Weaver and Company, Aurora-Colorado)

EEG conductive paste. In the normally hearing subjects the impedances of electrode pairs were checked and re-applied until they were $\leq 4 \text{ k}\Omega$ (37 Hz nominal). In the hearing-impaired subjects a less stringent impedance criteria of $\leq 8 \text{ k}\Omega$ (37 Hz nominal) was applied as the subjects were older (mean 70 years compared to mean 26) and had poorer skin condition which in some cases led to difficulties in achieving the lower impedance without causing skin irritation or discomfort. The recordings took place in a darkened sound-treated booth and the subjects were asked to relax and, if possible, sleep to reduce background EEG-noise levels and reduce any fluctuations in amplitude due to changes in attention (Cohen et al. 1991).

In the normally hearing subjects the signal level was set to 50 dB above the subjects' behavioral pure tone thresholds which were converted to dB SPL using the HL to SPL conversion for the coupler and insert earphone combination (BSI 1997).

In the hearing-impaired subjects the signal level was set to 20 dB above their ASSR-stimulus behavioral threshold (dB SL) as the amplitude intensity function for normally hearing and hearing-impaired subjects shows rapid growth in response amplitude once ASSR threshold is exceeded in hearing-impaired subjects (Picton 2010). Therefore lower signal sensation levels are required in hearing-impaired compared to normally hearing subjects to produce similar response amplitude levels. However, in some cases the response amplitudes recorded from a 20 dB SL stimulus were not significantly greater than the background EEG noise (2% F-test level). It was therefore, in those cases, necessary to repeat the recording at 20 dB SL within the same test session to determine if the response amplitude in a second recording reached the 2% significance level. In cases where 2% significance was not reached in this second recording in the first test session the signal level was increased to ensure that the response amplitude differences between sessions were valid and not solely a recording of EEG noise. This situation occurred for subject M5 whose recordings were performed using a higher (25 dB SL) signal level. In the second session, repeated recordings with increased signal level were not performed since the recording level was already established in the first session.

The ASSR system has the facility to reject (and repeat) epochs of the EEG that contained peak amplitudes higher than a criterion level. The EEG rejection level was determined in each recording session. The recording commenced using a $20 \mu\text{V}$ rejection level. If the number of rejections was 20% or less during the first four sweeps the recording was allowed to continue. If more than 20% of epochs were rejected during the first four sweeps the rejection level was increased in $5 \mu\text{V}$ steps until less than 20% of the epochs were rejected and the initial recording was stopped and discarded. The voltages across the vertex and neck electrodes were digitally sampled and recorded using the MASTER ASSR system with a Grass LP511 AC amplifier (Grass Technologies, Astro-Med, Inc. Product Group, Rhode Island, USA). The EEG amplifier was set to 10,000 times amplification with a high-pass filter of 30 Hz and low-pass filter of 300 Hz. The EEG sampling rate was set to 1000 Hz and 1000 samples per epoch were recorded. 15 sweeps containing 16-one second epochs were recorded in each session taking 4 minutes of test time plus additional test time for epochs repeated due to noise rejection. Stimulus artifact generated by aliasing of the voltages elicited on the EEG wires from the insert earphones was avoided by the use of -24 dB/octave 300 Hz low-pass filtering in the LP511 EEG amplifier (Picton et al. 2004). No stimulus artifact was recognized in a no-subject test condition, with the EEG inputs connected through a closed circuit resistor array representing the electrode impedances, with a sound output level of 120 dB SPL. The raw EEG data were saved and analyzed in MATLAB.

Data Analysis

The raw EEG sample data were imported into MATLAB for offline analysis. The EEG samples were band-pass filtered using a zero-phase filter with a 12 dB per octave slope with pass-band frequency range of 5 Hz above and below each analysis frequency and averaged using variance weighting (John et al. 2001; Picton et al. 2003). The averaging process for each individual test was performed fifteen times. The first analysis contained one sweep and in each of the fourteen subsequent analysis points the number of sweeps included in the averaging was incremented by one sweep. The averaged response waveform for each analysis time point was 16 seconds long (and contained 16,000 samples). The waveforms were converted into their frequency components using a Fast Fourier Transform (FFT) with 16 seconds of data giving a spectral resolution of 0.0625-Hz frequency bins. The energy (amplitude and phase) in the 60 frequency bins above and below the modulation-frequency bin were vector averaged to calculate the EEG-noise floor at each analysis point. In each analysis the energy in the modulation-frequency bin was compared for significance against the EEG-noise floor using the F-test technique that determines if the response is significantly different from the background EEG noise. Response latencies were calculated using phase to latency conversion (John et al. 2000). The multiple analysis averaging points enabled the response amplitude, phase, EEG noise and F-test significance level to be analyzed with increasing test time within each of the fifteen-sweeps recordings. Analyses were performed for each recording examining the amplitude, noise and F-test significance at the stimulus modulation frequency (95 Hz) and an arbitrarily chosen frequency point within the 80-Hz ASSR range (87 Hz). The EEG waveforms were filtered separately for each analysis modulation frequency with a separate filter with pass-band of 5 Hz above and 5 Hz below the analysis frequency (90 to 100 Hz pass-band for 95 Hz analysis and 82 to 92 Hz pass-band for 87 Hz analysis). The two analysis frequencies were chosen to consider both true and false response detection in a similar manner to the analysis performed by Luts et al. (2008).

RESULTS

The distributions of the evoked response amplitudes were checked for normality using the Shapiro-Wilks test and summarized using means and standard deviations.

It was necessary to increase the rejection level in 44 of the 60 recordings as in many cases 100% of epochs were rejected at the 20 μ V rejection level. The mean rejection levels were 31 μ V (SD 11) and 23 μ V (SD 4) for the normally hearing and hearing-impaired groups, respectively.

The recorded response and EEG-noise amplitude of the two fifteen-sweep recordings for each subject are summarized in Table 1. In subject N1's second recording session the response amplitude after 15 sweeps of recording time did not exceed the EEG-noise floor sufficiently to satisfy the F-test significance at the 1% level but was significant at the 2% level (see Table 2 and Figure 2).

In subject S2's second test run, the response amplitude reached the 2% significance level after 3 sweeps but was not significant (up to the 5% level), at the end of the 15 sweep test run. Figure 1 shows the response amplitude, the response amplitudes required to achieve 5% and 1% F-test significances, and the EEG-noise floor plotted against the number of sweeps included in the analysis for this test run. This plot shows the point in the recording where the response temporarily became significant then subsequently reverted to insignificant as test time increased. In subject M1, the first recording at 20 dB SL failed to produce a significant ASSR (2% level) at the end of the 15 sweep test run, but a second repeated recording in the first session did produce a significant ASSR. In the second session of subject M1, the

recording at 20 dB SL failed to produce response amplitude significantly larger than the EEG-noise floor (2% level) at the end of the 15 sweep test run. The two recordings, where significance was not reached at the end of the 15 sweeps in the second test sessions, were included in the analysis, because the responses in the first test session (after 15 sweeps) were significant at the 2% F-test level as required by our protocol.

Figures 2 and 3 show the significance level obtained at each time point (cumulative number of sweeps) in the twenty normally hearing subjects (two recordings per subject) and in the ten hearing-impaired subjects (two recordings per subject), respectively. Each shaded division of the figures represents the significance level in that recording (analyzed by averaging the EEG data including the number of sweeps indicated by the x-axis, representing an offline running average). Individual tests are shown across each row, with two rows per subject. The lower row for each subject is the first recording session, and the upper the second recording session. The left panel of the plot represents the recordings analyzed at 87 Hz to examine false response detection. The right panel shows the recordings analyzed at the modulation frequency of the ASSR stimulus (95 Hz). The x-axis depicts increasing test time. For example, significance analyses in column four (x-axis 4 sweeps) include sweeps one to four, significance analyses in column ten (x-axis 10 sweeps) include sweeps one to ten (16 seconds to 160 seconds).

In some cases, such as in normally hearing subjects N9 and N10, the 1% significance level was achieved after one sweep. In some recordings the significance levels fluctuated. For example, in the first recording session for subject N14, 5% significance was reached after 6 sweeps. Then, following this, the response was no longer significant as subsequent sweeps were added to the averaging, until after 10 sweeps, after which the response reached the 5% significance level. The response then reached, and maintained, the higher 1% significance level after the 13th sweep was included in the analysis.

False Response Detection

At 87 Hz the EEG energy should never have been significantly higher than the surrounding EEG noise (within the limits of the statistical test) as the stimulus was not modulated at 87 Hz. However, as shown in Figures 2 and 3 (left panels), using a test method in which testing stopped when significance reached the 5% F-test criterion level would have led to an incorrectly detected response in thirteen of the forty normally hearing subject recordings and two of the twenty hearing-impaired subject recordings. The error rates for the fixed test time (15 sweep) for the 5% F-test criterion were four out of 40 recordings for the normally hearing subjects and zero out of the twenty recordings for the hearing-impaired subjects. The results indicate that the false positive rate, shown as a percentage of the recordings, would be elevated from 6.7% for the ‘stop after 15 sweeps’ method to 25% when using the ‘stop testing when significance reaches the 5% criteria level’ method.

In three recordings the 2% significance level was reached, but not maintained with increasing test time (N3 first session, N20 first session, M5 second session). In two of the recordings (N14 first session, N10 second session) the 2% significance criterion level was reached and remained until the end of the recording (15 sweeps). The false positive rate at the 2% significance level using the “stop when significant” method, as a percentage of recordings, would have been 8.3% (5 out of 60) compared to 3.3% (2 out of 60) using the full test time (15 sweeps) protocol.

In order to examine the error rates further we considered protocols that could eliminate the elevated error rates found using the “stop when criterion significance level reached” method. Further analyses of the data presented in Figures 2 and 3 revealed that the increased error rates found when using the “stop when significant” protocol (5% level) could be eliminated

by requiring a 5% significance after minimum of 9 sweeps (v5-min9 protocol), or by requiring that the 5% significance remains over four consecutive sweep increments (v5-min4c protocol). These two additional protocols, referred to as v5-min9 and v5-min4c, were similar to the V4 (four consecutive), and V1-min8 (one consecutive, minimum 8 sweeps) protocols used by Luts et al. (2008).

Pairwise McNemar's binomial tests were performed between pairs of variable test time detection paradigms (v5-min4c, v5-min9, 5%, 2% and 1%) performed, see Table 3.

Detection time

The test time (number of sweeps) required to detect the responses using different detection protocols are shown in Table 2. The calculated detection times are the offline analysis detection times, which are lower than the true recording detections times. The offline analysis does not and cannot consider the time points of the rejected epochs throughout the recording as the MASTER ASSR system does not allow the time points of rejected epochs to be stored. The mean (standard deviation in parenthesis) number of sweeps required to detect responses and the number of false response detections for each detection protocol are summarized in Table 4. There were significant differences in the detection times between the five protocols (Friedman's test, $p < 0.05$, significance remained for all Friedman's pairwise comparisons). The v5-min4c protocol detected the response in a shorter test time compared to the v5-min9 protocol in 43 of the 58 recordings that reached significance in both protocols. The 2% "stop when significant" protocol detected the response in a shorter test time compared to the v5-min4c and the v5-min9 protocols in 51 out of the 58 recordings that reached significance in all three protocols.

Repeatability of response

The ASSR amplitudes and the EEG-noise levels in the adjacent FFT-frequency bins recorded from the two sessions are shown in Figure 4 for the normally hearing and hearing-impaired subjects. The noise corrected amplitudes were calculated using the equation: $\text{ASSR amplitude} = (\text{response}^2 - \text{noise}^2)$, where response is the amplitude of ASSR at the modulation frequency and noise is the mean EEG noise in the 60 FFT bins above and below the response amplitude used to calculate the F-test (Menard et al. 2008). The mean noise corrected amplitudes across all subjects and recording sessions were 73 nV and 59 nV for the normally hearing and hearing-impaired subjects, respectively. The spreads of individual subject mean amplitudes in the normally hearing and hearing-impaired groups were similar (standard deviation 24 nV for normally hearing group and 27 nV for hearing-impaired group). There was no significant difference in mean noise-corrected response amplitudes between the normally hearing and hearing-impaired groups ($t = 1.484$ $p = 0.149$).

Bland et al. (1986) proposed a method for assessing test repeatability by plotting the mean value of two recordings against the difference in value between the two recordings in groups of subjects. Their method quantifies test repeatability by estimating the test repeatability coefficient, which is calculated from twice the standard deviations of the differences. Therefore, approximately 95% of the differences fall between zero $\pm$ the repeatability coefficient (assuming an unbiased mean difference of zero). The mean amplitudes and amplitude differences between test runs were calculated from noise corrected amplitudes of each session to facilitate the Bland et al. (1986) repeatability analysis method. The mean amplitude and amplitude difference between recording sessions and the mean latency and latency difference between recording sessions are presented in plots as suggested by Bland et al. (1986) in Figures 5 and 6. The calculated repeatability coefficients for the ASSR amplitudes were 29 nV for the normally hearing subjects and 57 nV for the hearing-impaired subjects. The repeatability coefficient for the hearing-impaired subjects excluding

subject S1 (outlier) was 37 nV. The calculated repeatability coefficients for the ASSR latencies were 1.10 ms for the normally hearing subjects and 1.19 ms for the hearing-impaired. The repeatability coefficient was calculated from each subject's mean amplitude and latency. Therefore, higher amplitudes or latencies are more repeatable (smaller differences) in a relative sense but constant in an absolute sense.

The repeatability of the EEG noise (60 FFT frequency bins above and below the response amplitude) was also examined. The repeatability coefficient of the noise was calculated in the same manner as the repeatability coefficient of response amplitude. The mean noise was 16 nV (SD 4) in the normally hearing subjects and 20 nV (SD 5) in the hearing-impaired subjects. The repeatability coefficients of the noise were 10 nV (normally hearing) and 22 nV (hearing impaired).

DISCUSSION

Effect of EEG-noise floor on ASSR

The United Kingdom National Health Service draft pediatric ASSR threshold protocol (Stevens et al. 2009) recommended a maximum test time of four minutes per test run and a noise rejection level of 10 μ V. The response amplitudes in this present work were recorded over four minutes of test time but higher noise rejection levels were allowed. It was not possible to record ASSRs at this lower rejection levels in our adult subjects as this would have led to all the EEG data being rejected. In 57 of the 60 recordings the average EEG noise in the 120 bins adjacent to the response modulation frequency, used for the F-test, exceeded the noise-floor limit of 10 nV recommended by the United Kingdom National Health Service draft pediatric ASSR threshold protocol (Stevens et al. 2009). In many cases the participants we tested were able to relax or sleep. Our data suggest that the 10 nV noise floor limit may not be realistically achievable in adults. In this study in 59 out of the 60 recordings noise levels fell below the 30 nV floor limit used by D'Haenens et al. (2008). However, the response amplitudes in several of the hearing-impaired and one of the normally hearing subjects were near to 30 nV and therefore would not have been detected as significantly different from the EEG noise had the noise conditions in those recordings been in the 20-30 nV range. Therefore the 30 nV limit is likely to be too high as noise at that level can be higher than the expected response amplitude and therefore could lead to raised ASSR thresholds compared to behavioral thresholds.

Response detection criteria and test time

Several experimenters, when recording ASSRs, have used a "stop when significant" protocol to reduce test time. A "stop when significant" protocol is recommended by the United Kingdom National Health Service draft ASSR protocol that suggests stopping a recording once the 2% significance level is reached with other experimenters using the 5% significance level. As discussed above, false detection rates higher than 5% are expected for repeated 5% level F-test used in variable test time analysis. The 5% F-test level must be adjusted to 2% for a fifteen sweep recording (assuming data is correlated at $r=0.75$). Our analysis suggests that test time could be significantly reduced, by using the stop when response reaches significant criteria. We found that, consistent with the predicted increase in error ratio of repeated testing, using the "stop when significant" protocol produced a high number of false responses (25% of recordings for 5% F-test). Luts et al. (2008) extensively examined false detection rates using multiple response detection paradigms at the 5% significance level and reported similar error rates (30%). In determining error rates Luts et al. (2008) analyzed the recordings at multiple modulation frequencies unrelated to the ASSR stimuli they used, whereas we examined only one frequency. We recorded responses from single AMEXP ASSR stimuli whereas Luts et al. (2008) and D'Haenens et al. (2010) used

multiple stimuli with mixed amplitude and frequency modulation in each recording. The noise-weighted averaging used by Luts et al. (2008) examined the noise over a wide range of frequencies (77 to 115 Hz, excluding analysis frequencies) as their recordings had multiple modulation frequencies over this wide frequency range. It is note-worthy that D'Haenens et al. (2010) investigated the false detection rate by analysing the EEG data at frequencies that could have included a true, weak response. The response was considered false when the “stop when significant” protocols showed a significant response whereas no response was detected when the full recording time was analyzed (64 sweeps). D'Haenens et al. (2010) method may have had increased false response detection rates compared to the true error rate because they assumed that the significance at the end of the recording was accurate and any detection of significant response with “stop when significant” protocols before the maximum test time was reached, was considered false. Since D'Haenens et al. analyzed their ASSRs for signals presented at 0 dB HL in normally hearing subjects, 11.3 % of their 248 responses were significant at the end of 64th sweep. Thus, it is possible that throughout each recording the noise and response may have fluctuated and the weak responses were detected at transient points throughout a recording. This could have occurred when the level of EEG noise was lower and the size of the response higher at points in the recording, and conversely the possibility of an insignificant response at the end (64th sweep) of the recording if the noise towards the end of the recording was higher and response lower. In contrast to this, Luts et al. (2008) and our present work considered false detection at modulation frequencies where no stimulus was presented eliminating this confounding factor. Overall, both the current results and those of Luts et al. and D'Haenens et al. are consistent in showing error rates higher than the F-test percentages used due to the increased chance of errors associated with repeated testing.

We also examined the false detection rate using the “stop when significant” protocol at the 2% F-test level, which was not performed by the previous experimenters, and found the false positive rate in our recordings was 8.3%. In our analysis there were four occurrences of false response detection using the “stop when significant” and one occurrence of false response detection using the “stop after 15 sweeps” (full test time), at the 1% F-test level. However, using the 1% level “stop when significant” prevented three of the responses, that were detected at the 2% level, from being detected within 15 sweeps as they never reached 1% significance (Subjects N1 run 2, M3 run 2 and M4 run 1). Luts et al. (2008) concluded that the 1% F-test in some cases prevented response detection and deemed it to be too strict.

Our analysis suggests that increased error rates found when using the “stop when significant” protocol at the 5% significance level compared to full fifteen sweep test time can be reduced by requiring significance to remain over four consecutively increasing number of sweeps (v5-min4c). However, both the v5-min4c protocols required greater test time compared to the 5% and 2% “stop when significant” protocols.

A greater number of false response detections occurred in the normally hearing group compared to the hearing-impaired group. It is possible that this difference was due to factors such as subject age, ability to relax (EEG-noise and rejection level) and electrode impedances, which are discussed in the following section.

Response Amplitude Repeatability

The estimated repeatability coefficients of the recorded response amplitudes between two test sessions indicate a high level of response amplitude variability with the repeatability coefficient being similar to that recorded by D'Haenens et al. (2008). In our case the variability was 40% of the mean amplitude in the normally hearing subjects (mean amplitude 73 nV, repeatability coefficient 29 nV) and 97% of the mean amplitude in the hearing-impaired (mean amplitude 59 nV, repeatability coefficient 57 nV), compared to

29% (mean 52 nV, ± 2 SEM repeatability 15 nV) for a 50 dB HL 2-kHz stimulus in normally hearing adults recorded by D'Haenens et al. (2008). The repeatability of the EEG-noise amplitudes were 13% and 37% of the mean response amplitude in the normally hearing and hearing-impaired subjects, respectively. The noise repeatability coefficients expressed as a percentage of mean response amplitudes are lower than the response amplitude repeatability coefficients expressed as a percentage of mean amplitudes suggesting that the differences in noise do not fully account for the differences in response amplitude. It is possible that the variable rejection levels used in this present work contributed to the variability of response and noise. However, it was not possible to use one universally applicable, low rejection level, as this would have led to excessively long recording times. Similarly, the use of one universal high rejection level would not have been optimal as it would have increased the overall noise in the recordings. This could equally affect variability, as the noise rejection would have been more or less effective between subjects and recordings depending on the maximum EEG-noise levels of each recording. Recorded residual EEG-noise can be reduced by decreasing EEG-noise rejection levels. However, the mean EEG-noise amplitudes were significantly lower in the normally hearing group compared to the hearing-impaired group ($t=-2.811$ $p=0.009$) but paradoxically the mean rejection level was higher in the normally hearing group compared to the hearing-impaired group. It is possible that other factors unrelated to the rejection levels, such as subject age or electrode impedances affected the EEG-noise (see discussion below). D'Haenens et al. (2008) found the response amplitude repeatability in normally hearing subjects, expressed as a percentage of mean response amplitude, to be poorer nearer to threshold than at higher levels. The response amplitude repeatability that we measured accounted for difference in EEG-noise amplitudes between sessions by using EEG-noise corrected response amplitudes whereas D'Haenens et al. (2008) compared raw response amplitudes. In our present work there were two other factors that could have affected the response amplitude repeatability found in each of the two groups (normally hearing and hearing-impaired). Firstly, the hearing-impaired subjects (mean age 70; SD 4 years) were older than the normally-hearing subjects (mean age 26; SD 7 years). At present it is not clear if age affects the variability of response amplitude either between or within recording sessions. However data of Purcell et al. (2004) and Petrescu and van Roon published in (Picton 2010) suggest that age might be a factor affecting the size and stability of the ASSR. Secondly, the hearing-impaired subject protocol had a less stringent impedance criterion. In 6 of the 20 recording sessions with hearing-impaired subjects, the impedances pairs exceeded the 4 k Ω impedance limit used with normally hearing subjects. The impedance pairs in all recordings of the hearing-impaired group were less than 8 k Ω as required by the protocol. The mean (SD) active-to-reference impedances were 2.5 k Ω (0.7) and 3.4 k Ω (1.6) in the normally hearing and hearing-impaired groups, respectively. There was no significant difference in the mean (active-to-reference) impedance between the normally hearing and hearing-impaired groups (mean of session 1 and session 2 impedance, $t=1.74$ $p=0.111$ equal variance not assumed as Levene's test $F=8.45$ $p=0.007$). Therefore there is no evidence to suggest that impedance was a confounding factor.

Our results show that 80-Hz-range amplitudes were highly variable between sessions but neither this present study nor D'Haenens et al. (2008) examined within-session amplitude variation. It is possible that variability of response amplitudes between sessions could have caused the ± 10 dB range of threshold repeatability (SEM method) of the ASSR threshold measurements found by D'Haenens et al. (2008). It is also possible that in their study high or differing levels of EEG-noise caused the threshold differences, but these data were not presented

CONCLUSIONS

The results of this study show that ASSR amplitudes are highly variable between test sessions with an average absolute estimated variability in response amplitude of 29 nV for normally hearing subjects, similar to D'Haenens et al. (2008) findings, and 57 nV for hearing-impaired subjects. This could be a possible cause of test retest differences in ASSR threshold measurements, as it could potentially lead to thresholds that were above the EEG-noise level and significant in one test session subsequently falling below the EEG-noise level in the repeat test session. The present study examined the repeatability of ASSR amplitudes at supra-threshold levels at one stimulus frequency and the confidence limits of the estimated repeatability coefficients were wide.

In 95% of our recording sessions the background EEG-noise levels were higher than the current draft United Kingdom National Health Service pediatric ASSR threshold protocol allows. We also found that 4 minutes of test time may be inadequate to detect supra-threshold responses in adults, depending on the EEG-noise conditions and chosen detection significance level. In the present study we cannot comment on the suitability of the draft protocol for use in infants as we did not test that age group. However, our findings suggest further work is needed to standardize recording parameters in adults, as the suggested pediatric parameters may not be suitable.

In all statistical tests false positives are expected at the level indicated by the chosen significance levels, which are increased in variable test time protocols due to repeated testing. Overall, the 5% significance level "stop when significant" response detection protocol required the least detection time, but this protocol also produced the highest number of false response detections. The high numbers of errors can be reduced by requiring the response to remain significant over four consecutively increasing sweep times or by using the 2% significance level (just below the 15-sweep adjusted 5% level). The 2% significance level "stop when significant" protocol, as suggested in the draft United Kingdom ASSR threshold protocol (Stevens et al. 2009), appears to be an appropriate trade-off between fast response detection time and minimizing false response detection. However, it is important to note that since our data were tested using multiple detection paradigms the results are likely to be impacted by errors associated with repeated testing (i.e incorrect false and true response detection rates). Nevertheless, this work further demonstrates the need for caution in applying variable test time protocols in ASSR detection. The impact of variable test time protocols have yet to be fully considered by all commercial ASSR systems.

Since we did not extensively test all frequencies and signal levels, we recommend further work to determine the most suitable ASSR parameters. Additionally, further work is required including a greater number of subjects and a wider range of signal levels and frequencies to further determine the repeatability of amplitudes, limits of EEG-noise floor, required test time, and appropriate response significance techniques, such as significance levels and variable test time criteria.

Acknowledgments

We would like to thank Michael Sasha John for his advice on technical issues and for lending us the MASTER equipment. The work was supported by a United Kingdom Medical Research Council PhD studentship and Medical Research Council Project Grant G0802190.

We thank two anonymous reviewers for their comments on the earlier version of this manuscript.

REFERENCES

- Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1:307–310. [PubMed: 2868172]
- BSI. BS EN ISO 389-2:1997 Acoustics - Reference zero for the calibration of audiometric equipment. Part 2: Reference equivalent threshold sound pressure levels for pure tones and insert earphones. British Standards Institute; 1997.
- Champlin CA. Method for detecting auditory steady-state potentials recorded from humans. *Hear Res*. 1992; 58:63–69. [PubMed: 1559907]
- Cohen LT, Rickards FW, Clark GM. A comparison of steady-state evoked potentials to modulated tones in awake and sleeping humans. *J Acoust Soc Am*. 1991; 90:2467–2479. [PubMed: 1774415]
- D’Haenens W, Vinck BM, De Vel E, et al. Auditory steady-state responses in normal hearing adults: a test-retest reliability study. *Int J Audiol*. 2008; 47:489–498. [PubMed: 18698523]
- D’Haenens W, Vinck BM, Maes L, et al. Determination and evaluation of clinically efficient stopping criteria for the multiple auditory steady-state response technique. *Clin Neurophysiol*. 2010; 121:1267–1278. [PubMed: 20457007]
- Dobie RA, Wilson MJ, Dobie RA, et al. A comparison of t test, F test, and coherence methods of detecting steady-state auditory-evoked potentials, distortion-product otoacoustic emissions, or other sinusoids. *Journal of the Acoustical Society of America*. 1996; 100:2236–2246. [PubMed: 8865632]
- Galampos R, Makeig S, Talmachoff PJ. A 40-Hz auditory potential recorded from the human scalp. *Proceedings of the National Academy of Sciences of the United States of America*. 1981; 78:2643–2647. [PubMed: 6941317]
- John MS, Dimitrijevic A, Picton TW. Weighted averaging of steady-state responses. *Clin Neurophysiol*. 2001; 112:555–562. [PubMed: 11222980]
- John MS, Dimitrijevic A, Picton TW. Auditory steady-state responses to exponential modulation envelopes. *Ear & Hearing*. 2002; 23:106–117. [PubMed: 11951847]
- John MS, Picton TW. Human auditory steady-state responses to amplitude-modulated tones: phase and latency measurements. *Hearing Research*. 2000; 141:57–79. [PubMed: 10713496]
- John, MS.; Purcell, D. Introduction to Technical Principles of Auditory Steady-State Response Testing. In: Rance, G., editor. *Auditory Steady-State Responses*. Plural Publishing; San Diego: 2008. p. 11–53.
- Kaf WA, Sabo DL, Durrant JD, et al. Reliability of electric response audiometry using 80 Hz auditory steady-state responses. *International Journal of Audiology*. 2006; 45:477–486. [PubMed: 17005490]
- Luts H, Van Dun B, Alaerts J, et al. The influence of the detection paradigm in recording auditory steady-state responses. *Ear Hear*. 2008; 29:638–650. [PubMed: 18469712]
- Menard M, Gallego S, Berger-Vachon C, et al. Relationship between loudness growth function and auditory steady-state response in normal-hearing subjects. *Hear Res*. 2008; 235:105–113. [PubMed: 18093768]
- Picton, TW. *Human Auditory Evoked Potentials*. Plural Publishing Inc; San Diego, USA: 2010.
- Picton TW, John MS. Avoiding electromagnetic artifacts when recording auditory steady-state responses. *Journal of the American Academy of Audiology*. 2004; 15:541–554. [PubMed: 15553654]
- Picton TW, John MS, Dimitrijevic A, et al. Human auditory steady-state responses. *International Journal of Audiology*. 2003; 42:177–219. [PubMed: 12790346]
- Purcell DW, John SM, Schneider BA, et al. Human temporal auditory acuity as assessed by envelope following responses. *J Acoust Soc Am*. 2004; 116:3581–3593. [PubMed: 15658709]
- Stevens, J.; Brennan, S.; Lightfoot, G. [Retrieved 28 July 2009] Provisional guidelines for using Auditory Steady State Responses (ASSR) in babies. 2009. from <http://hearing.screening.nhs.uk/cms.php?folder=4113>
- Sturzebecher E, Cebulla M, Elberling C. Automated auditory response detection: statistical problems with repeated testing. *Int J Audiol*. 2005; 44:110–117. [PubMed: 15913159]

- Valdes JL, Perez-Abalo MC, Martin V, et al. Comparison of statistical indicators for the automatic detection of 80 Hz auditory steady state responses. *Ear & Hearing*. 1997; 18:420–429. [PubMed: 9360865]
- Wilding T, McKay C, Baker R, et al. Using the Auditory steady-state response to record response amplitude curves. A possible fast objective method for diagnosing dead regions. *Ear & Hearing*. 2011 In press.

Fig 1.

Response amplitude (darker thicker line), significant response amplitude (1% criterion) level (lighter dashed and dotted line), significant response amplitude (5% criterion) level (darker dashed and dotted line) and noise (dashed line) against number of consecutive sweeps in the analysis for the second test session of subject S2. The point at which the response reached the 5% significance level is marked with an arrow head.

Fig 2.

F-test significance levels shown against test time (number of sweeps) for forty recordings in normally hearing subjects at 50 dB SL (two recordings for twenty subjects). Individual shaded rectangles represent the calculated significance level at each position in the recording. Rows represent individual recordings with two recordings for each subject. Columns represent the number of sweeps included in the response calculation. The left panel shows the false responses detected at 87 Hz and the right panel shows true response detection at the stimulus modulation frequency of 95 Hz. White shading indicates response amplitudes not significant (5% level), light grey significant between 2% and 5%, dark grey significant between 2% and 1%, and black significant at <1%.

Fig 3.

As for figure 3 except for twenty recordings in hearing-impaired subjects (two recordings per subject).

Fig 4. Response amplitudes recorded from normally hearing (N1-N20 upper panel) and hearing-impaired subjects (M1-M5 and S1-S5 lower panel).

Fig 5.

Repeatability of ASSR amplitudes in normally hearing subjects (N1-N20) in the left panel and hearing-impaired subjects (M1-M5 and S1-S5) in the right panel. Data points show the mean response amplitude (nV on the x-axis) vs. the difference in response amplitude between two test runs (nV on the y-axis). The upper and lower dashed lines show the assumed mean difference of $0 \text{ nV} \pm$ the calculated repeatability coefficient. The error bars show the 95% confidence limits of the repeatability coefficients.

Fig 6.

As for figure 5 except for mean response latency (x-axis) and difference in response latency between two test runs (y-axis).

TABLE 1

ASSR recording analysis for normally hearing subjects (N), mild hearing loss subjects (M) and severe hearing loss subjects (S) in session 1 and session 2. F-ratios marked in bold indicate cases where the response amplitude was not significantly above the EEG-noise level (2% level). In subject S2 the 2nd recording reached significance after 3 sweeps but was not significant after 15 sweeps.

Subject	Session 1				Session 2			
	Amplitude (nV)	Latency (mS)	Noise (nV)	F ratio p	Amplitude (nV)	Latency (mS)	Noise (nV)	F ratio p
N1	55	16.5	14	0.000	47	15.6	24	0.019
N2	126	16.8	26	0.000	116	17.4	18	0.000
N3	50	18.0	15	0.000	43	17.5	17	0.002
N4	68	18.1	14	0.000	96	17.7	18	0.000
N5	98	17.7	14	0.000	78	17.1	12	0.000
N6	68	16.4	12	0.000	85	16.9	13	0.000
N7	60	16.3	12	0.000	72	15.6	24	0.000
N8	61	16.2	13	0.000	49	15.6	12	0.000
N9	120	16.8	11	0.000	124	16.9	12	0.000
N10	56	16.9	11	0.000	54	17.9	13	0.000
N11	50	16.1	17	0.000	65	15.6	11	0.000
N12	102	16.1	13	0.000	110	16.0	15	0.000
N13	55	16.6	11	0.000	65	16.0	21	0.000
N14	55	17.2	20	0.001	61	17.0	21	0.000
N15	94	17.6	22	0.000	117	17.2	24	0.000
N16	74	16.1	14	0.000	97	16.4	13	0.000
N17	37	15.6	11	0.000	59	15.2	10	0.000
N18	83	15.9	13	0.000	77	16.1	16	0.000
N19	76	15.6	14	0.000	83	15.8	15	0.000
N20	64	14.8	16	0.000	54	15.7	18	0.000
Mean(SD)	73(24)	16.7(0.8)	15(4)		78(25)	16.5(0.8)	16(4)	
M1	35	13.5	8	0.000	48	14.5	32	0.112
M2	64	16.1	23	0.001	56	17.0	20	0.001

Subject	Session 1				Session 2			
	Amplitude (nV)	Latency (mS)	Noise (nV)	F ratio p	Amplitude (nV)	Latency (mS)	Noise (nV)	F ratio p
M3	69	16.2	27	0.002	42	15.3	21	0.016
M4	34	15.4	17	0.015	50	15.9	18	0.001
M5	64	17.1	26	0.003	81	17.5	23	0.000
Mean(SD)	56(23)	14.3(4.1)	17(1)		67(31)	15.7(1.7)	22(7)	
S1	76	17.9	19	0.000	145	17.8	14	0.000
S2	33	15.7	6	0.000	36	16.3	29	0.220
S3	38	13.8	11	0.000	49	13.7	9	0.000
S4	106	16.6	19	0.000	85	16.4	23	0.000
S5	59	12.4	23	0.002	88	12.4	26	0.000
Mean (SD)	62(30)	15.3(2.2)	16(7)		81(43)	15.3(2.2)	20(8)	

TABLE 2

ASSR recording analysis for normally hearing subjects (N), mild hearing loss subjects (M) and severe hearing loss subjects (S) in session 1 and session 2. Four “stop when significant” detection protocols were applied. Conditions one, two and three: “1%”, “2%” and “5%” indicate the number of sweeps for the recording to reach the required significance level. Condition four “5% > 9 sweeps”, indicates a minimum of 9 sweeps. Condition five: “5% level > 4 consecutive sweeps” indicates significance remains as 4 consecutive sweeps are added to the analysis. Empty data points marked as “-” indicate that significance was not reached for that condition.

Subject	Session 1 Sweeps to significance					Session 2 Sweeps to significance				
	1% level	2% level	5% level	5% level > 9 sweeps	5% level > 4 consecutive sweeps	1% level	2% level	5% level	5% level > 9 sweeps	5% level > 4 consecutive sweeps
N1	6	6	5	9	8	-	14	13	13	-
N2	3	2	2	9	5	1	1	1	9	4
N3	11	5	5	10	13	10	9	6	9	12
N4	6	4	4	9	7	6	6	4	9	9
N5	3	2	1	9	4	2	2	2	9	5
N6	2	2	2	9	5	1	1	1	9	4
N7	2	2	1	9	4	2	2	1	9	4
N8	4	2	2	9	5	5	3	3	9	8
N9	1	1	1	9	4	1	1	1	9	4
N10	1	1	1	9	4	4	3	1	9	6
N11	11	9	8	9	11	3	3	3	9	6
N12	2	2	2	9	5	2	1	1	9	4
N13	4	4	3	9	6	10	9	7	9	10
N14	14	13	6	11	14	11	11	11	11	14
N15	4	4	3	9	6	2	2	2	9	5
N16	4	4	1	9	6	1	1	1	9	4
N17	7	2	2	9	9	2	2	2	9	5
N18	1	1	1	9	4	3	3	1	9	6
N19	5	5	4	9	7	5	5	4	9	7
N20	6	6	5	9	8	9	8	8	9	11
M1	2	1	1	9	4	1	1	1	9	11

Subject	Session 1 Sweeps to significance					Session 2 Sweeps to significance				
	1% level	2% level	5% level	5% level > 9 sweeps	5% level > 4 consecutive sweeps	1% level	2% level	5% level	5% level > 9 sweeps	5% level > 4 consecutive sweeps
M2	3	3	3	9	6	2	2	1	9	4
M3	10	8	7	9	10	12	11	8	10	13
M4	11	5	3	9	6	10	4	4	9	7
M5	12	10	9	9	12	12	11	8	9	11
S1	5	5	5	9	8	1	1	1	9	4
S2	5	5	2	9	8	-	3	3	-	0
S3	5	5	2	9	5	1	1	1	9	4
S4	1	1	1	9	4	4	4	1	9	7
S5	10	5	4	9	11	4	4	3	9	6

Table 3

Pairwise McNemar binomial test p values for variable test time response detection paradigms. $p < 0.005$ marked in bold.

Detection paradigm	2%	5%	V5-min4c	V5-min9
1%	1.000	0.001	1.000	0.250
2%	-	0.002	1.000	0.625
5%	0.002	-	0.001	0.008

TABLE 4

Mean response detection times and number of false positive response detections for 60 recordings. False response detections analysed at 87 Hz, true responses at 95 Hz.

		Variable test time					Fixed test time (15 sweeps)		
		1%	2%	5%	v5-min9	v5-min4c	1%	2%	5%
Normally Hearing	Mean, median (SD) sweeps	4.5 4.0 (3.5)	4.1 3.0 (3.4)	3.3 2.0 (2.9)	9.2 9.0 (0.8)	6.7 6.0 (3.0)	>15	15	15
	False response detections (% of recordings)	4.0 (10.0)	4.0 (10.0)	13.0 (32.5)	7.0 (17.5)	4.0 (10.0)	1 (2.5)	2 (5.00)	4 (10.00)
Hearing Impaired	Mean, median (SD) sweeps	6.2 5.0 (4.3)	4.5 4.0 (3.3)	3.4 3.0 (2.7)	9.1 9.0 (0.2)	7.1 6.5 (3.4)	>15	>15	>15
	False response detections (% of recordings)	0 (0)	1.0 (5.0)	2.0 (10.0)	0 (0)	0 (0)	0	0	0
All	Mean, median (SD) sweeps	5.0 4.0 (3.8)	4.2 3.0 (3.3)	3.3 2.0 (2.8)	9.2 9.0 (0.4)	6.8 6.0 (3.0)	>15	>15	>15
	False response detections (% of recordings)	4.0 (6.7)	5.0 (8.3)	15.0 (25.0)	7.0 (11.7)	4.0 (6.7)	1 (1.7)	2 (3.3)	4 (6.7)