The Convolutional Group Sequential Test: reducing test time for evoked potentials

Michael A. Chesnaye, Steven L. Bell, James M. Harte, & David M. Simpson

***Abstract*—When using a statistical test for automatically detecting evoked potentials, then the number of stimuli presented to the subject (the sample size for the statistical test) should be specified at the outset. For evoked response detection, this may be inefficient, i.e. because the signal-to-noise ratio (SNR) of the response is not known in advance, the user would usually err on the cautious side and use a relatively high number of stimuli to ensure adequate statistical power. A more efficient approach is to apply the statistical test repeatedly to the accumulating data over time, as this allows the test to be stopped early for the high SNR responses (thus reducing test time), or later for the low SNR responses. The caveat is that the critical decision boundaries for rejecting the null hypothesis need to be adjusted if the intended type-I error rate is to be obtained. This study presents an intuitive and flexible method for controlling the type-I error rate for sequentially applied statistical tests. The method is built around the discrete convolution of truncated probability density functions, which allows the null distribution for the test statistic to be constructed at each stage of the sequential analysis. Because the null distribution remains tractable, the procedure for finding the stage-wise critical decision boundaries is greatly simplified. The method also permits data-driven adaptations (using data from previous stages) to both the sample size and the statistical test, which offers new opportunities to speed up testing for evoked response detection.**

***Index Terms*—sequential testing, evoked potentials, objective detection methods, data-driven adaptations**

I. Introduction

Evoked potentials are changes in neurophysiological activity within the peripheral or central nervous system, time-locked to externally applied sensory stimuli [9]. They can be recorded invasively, or non-invasively using electroencephalography (EEG), and can be used to: (i) demonstrate or confirm an abnormal functioning of the sensory or central nervous system, (ii) explore underlying anatomical structures, (iii) provide insight into pathophysiology, and (iv) monitor changes in neurological activity, e.g. for intra-operative monitoring [27]. For many of these applications, the first step is to determine whether a response is present or not, which can be achieved objectively by applying a statistical test to the acquired data and generating a p value, i.e. a probability that the null hypothesis, H_0, of 'no evoked response present' is true. Such statistical tests avoid the need for highly trained specialists, who are otherwise given the task to manually inspect the acquired data, along with the associated experimenter-dependent subjective judgements. There are many different statistical tests available for response detection, e.g. the FSP [10], various Q-sample statistics [3], the Magnitude Squared Coherence [24], bootstrapped statistics [19], and the Hotellings T^2 test [5, 12], just to name a few. When using a conventional approach, the statistical test is applied to the data just once after all data has been collected, henceforth referred to as a 'single shot' test. This may be inefficient in terms of data needed to accept or reject H_0.

In general, the main challenge in detecting evoked potentials within the ongoing EEG activity is their low signal-to-noise ratios (SNRs, defined as the power of the response relative to background noise). The auditory brainstem response (ABR), for example, has a peak amplitude of around 0.5 μV [14], whereas the EEG background activity can have amplitudes in the range of at least 10 μV after filtering. Many stimuli therefore need to be presented to the subject, and the resulting EEG averaged to reduce residual noise levels, before an unambiguous response can be detected. The SNR of the evoked response can furthermore vary significantly both between and within recordings due to non-stationary EEG background activity (varying noise levels in the recording environment), changes in the acoustic stimulus, variations in response amplitude and response morphology, or varying electrode impedances. Note therefore that any *a priori* choice for the sample size will tend to result in either an over-powered test (and an unnecessarily prolonged test time) for the higher SNR responses, or an under-powered test (and potentially an increased type-II error rate, i.e. a reduced test sensitivity) for the lower SNR responses.

A solution to uncertainty in the SNR is to apply a number of statistical tests sequentially to the accumulating data over time. This allows the test to be stopped early for the higher SNR responses, thus reducing test time, or later for the lower SNR responses. The challenge with such sequential test procedures is that the probability of incorrectly rejecting H_0 is increased with the number of interim looks at the data. The latter is also known as an 'inflated' type-I error rate, and adjusted critical decision boundaries, for rejecting or accepting H_0, are required if the desired type-I error rate is to be obtained.

The main aim for this paper is to present a simple and intuitive method for finding the critical decision boundaries and controlling the type-I error rate of sequentially applied statistical tests. The method, called the Convolutional Group Sequential Test (or CGST), is similar to previous methods [1, 2, 4, 14, 26] in that data is analysed incrementally, in disjoint groups of samples. At each stage of the sequential analysis, a group of samples is analysed with a statistical test, and a p value is generated. The null hypothesis is then evaluated using a summary statistic, composed of all stage-wise p values. The goal is hence to construct the null distribution for this sum-

mary statistic, achieved by numerically convolving truncated probability density functions. Because the null distribution for the summary statistic remains tractable, the procedure for finding the stage-wise critical decision boundaries is greatly simplified. The main advantage over some alternative sequential test procedures is flexibility, simplicity (including low computational load) and clarity in interpretation.

The remainder of this paper is structured as follows: the underlying theoretical framework of the CGST is first described in section II, after which some simulation results are presented in section III. The goal for the simulations is to explore the performance of the CGST across a range of SNRs and CGST design parameters. An analysis of subject-recorded ABR data is then also presented, with the goal to provide an illustrative example, and to further demonstrate the benefits of using a sequentially applied statistical test in practice. In section IV, various trade-offs associated with CGST design parameters are discussed, and the adaptive group sequential test is considered in more detail. Various connections between the CGST and existing methods are also drawn.

II. Theoretical framework and graphical illustrations

This section introduces the notation and underlying theoretical framework for the CGST, after which graphical illustrations are used to further clarify the approach. Consider first a sequential test procedure with K stages, i.e. the statistical test is applied to the data K times, with each stage considering a new group of independent samples. The choice for the statistical test will depend on the specific problem, but does not affect the CGST itself. The goal is to evaluate the global null hypothesis H_0 at nominal significance level α:

$$H_0 : H_{01} \cap ... \cap H_{0K} \tag{1}$$

where H_{0i} (for $i = 1, 2, ..., K$) is the null hypothesis at stage i. In the current work, all stage-wise null hypotheses H_{0i} are defined as 'evoked response not present'. At each stage, a new group of samples is collected, and a p value is generated by analysing this group of samples with a statistical test (e.g. the Hotellings T2 test, Q-samples, MSC or any other, as only the p value is required). Similar to [1, 2, 4, 16, 26], it is assumed that all stage-wise p values p_i (for $i = 1, 2, ..., K$) are stochastically independent, which implies that the accumulated evoked response data cannot be pooled, but must be analysed in disjoint sub-samples. Data analysed in stage i, for example, cannot be re-analysed in the subsequent stages of the trial, neither can it be pooled with previously collected data. However, at each stage of the analysis, all stage-wise p values can be combined into a summary statistic, after which the test can be stopped for either futility or efficacy, or the test proceeds to the next stage. Futility implies that the summary statistic is sufficiently far from statistical significance, such that additional data collection is deemed futile, and H_0 is accepted, whereas efficacy implies that there is sufficient evidence for rejecting H_0 at level α. The CGST furthermore requires the summary statistic to be a summation of the (potentially transformed) p values. The stage k summary statistic is thus defined as:

$$\Sigma_k = \sum_{i=1}^{k} f_i(p_i) \tag{2}$$

where $f_i(p_i)$ is the desired transformation at stage i for p_i. A typical transformation that may be used here is that of Fisher [11], achieved by defining $f_i(p_i) = -2\ln(p_i)$. When p_i is uniform on [0,1] under H_0, then $-2\ln(p_i)$ is χ^2_2-distributed. Note that although transformation is not necessary, combining the original p values through summation can potentially result in a small loss of test sensitivity relative to some alternative combination functions (see e.g. [7]). Fishers method in particular has some desirable properties in terms of efficiency [18], which can be attributed to the $\ln(p_i)$ transform placing more emphasis on small p values, and because a succession of small p_i is more likely when an evoked response is present.

After combining the stage-wise p values, the test can be stopped at stage i for futility when $\Sigma_i < C_i$, or for efficacy when $\Sigma_i > A_i$, where A_i and C_i (for $i = 1, 2, ..., K$) are the stage i critical decision boundaries. Note that it is assumed here that transformation $f_i(p_i)$ gives large values for small p_i, i.e. that $f_i(p_i)$ is monotonic with a negative gradient.

Critical decision boundaries

The method for finding the critical decision boundaries A_i and C_i, such that the nominal α-level of the full test is preserved, is built around the convolution theorem, which states [13]:

> *The null distribution for the sum of two independent random variables is given by the convolution of their individual null distributions.*

Hence, if the stage-wise null distributions (the null distributions for $f_i(p_i)$, henceforth denoted by ϕ_i) are known, then these can be iteratively convolved to find the null distribution for the combined statistic Σ_i, henceforth denoted by ϕ_{Σ_i}. An important caveat is that ϕ_{Σ_i} changes when proceeding from stage $i-1$ to stage i, as it is not possible to enter stage i with $\Sigma_{i-1} > A_{i-1}$ or $\Sigma_{i-1} < B_{i-1}$, else the trial would already have been stopped. The H_0 rejection and acceptance regions for $\phi_{\Sigma_{i-1}}$ should therefore be truncated prior to convolving with ϕ_i. More formally, the null distribution for the combined statistic at stage two is given by:

$$\phi_{\Sigma_2} = \phi_1^{T[C_1, A_1]} * \phi_2 \tag{3}$$

and for all following stages by:

$$\phi_{\Sigma_i} = \phi_{\Sigma_{i-1}}^{T[C_{i-1}, A_{i-1}]} * \phi_i \tag{4}$$

where $*$ denotes convolution, and where $\phi^{T[C,A]}$ indicates that distribution ϕ contains non-zero values exclusively for the $[C, A]$ interval (the distribution has been truncated to this interval).

Once ϕ_{Σ_i} has been generated, then finding A_i and C_i is straightforward. In particular, the stage i critical boundary for efficacy, A_i, is found by numerically solving:

$$\Phi_{\Sigma_i}[A_i, \infty] = \alpha_i \tag{5}$$

where α_i is the desired type-I error rate for stage i, and where $\Phi_{\Sigma_i}[A_i, \infty]$ is the cumulative distribution function for Σ_i, calculated across the interval $[A_i, \infty]$. In practice, ∞ is of course replaced by a sufficiently large value. The α_i values (for $i = 1, 2, ..., K$) are furthermore chosen freely, under the condition that $\sum_{i=1}^{K} \alpha_i = \alpha$. Similarly, the stage i critical boundary for futility, C_i, is found by numerically solving:

$$\Phi_{\Sigma_i}[0, C_i] = \gamma_i \tag{6}$$

where γ_i is the stage i fraction of tests to be rejected for futility when H_0 is indeed true, i.e. the stage i true-negative rate (TNR). The γ_i values (for $i = 1, 2, ..., K$) are also chosen freely, under the condition that $\alpha + \sum_{i=1}^{K} \gamma_i \leqslant 1$.

It is worth emphasizing here that ϕ_{Σ_i} can be viewed as the joint PDF of variables $f_i(p_i)$ and $\Sigma_{i-1}^{T[C,A]}$, where $\Sigma_{i-1}^{T[C,A]}$ is the stage $i-1$ summary statistic, restricted to the specific range of values (the [C,A] interval) for which entry to stage i is ensured. Note also that the area under the joint PDF ϕ_{Σ_i} is decreased with each additional truncation, which implies a limit to the total number of stages permitted. In particular, the total area under ϕ_{Σ_k}, say AR_k, after the stage $k-1$ truncations is given by $AR_k = 1 - \sum_{i=1}^{k-1} \gamma_i + \alpha_i$.

For the remainder of this paper, it is assumed that all stage-wise p values are uniformly distributed on [0,1] under H_0 ($\phi_i \sim U(0,1)$ for all i), as is customary for evoked response detection. This assumption is valid when H_0 is true and the assumptions underlying the statistical test are satisfied. If assumptions are violated (typically due to non-stationarity of the EEG signals, non-Gaussianity of the data, or serial correlation between epochs), then the distribution of p values will be non-uniform, and the critical decision boundaries generated by the CGST will be inaccurate.

Graphical Illustrations

The goal for this section is to clarify the procedure using graphical illustrations and a generic example. First, let the nominal α-level be 0.15 (an unusually high type-I error rate is chosen for illustration purposes only), and be spread equally across 3 stages ($K = 3$), giving stage-wise type-I error rates $\alpha_1 = \alpha_2 = \alpha_3 = 0.05$. The γ_i values are furthermore specified as $\gamma_1 = 0.2$, $\gamma_2 = 0.4$, and $\gamma_3 = 0.25$, such that $\alpha + \sum_{i=1} \gamma_i = 1$ (further considerations on how to choose the stage-wise α_i and γ_i values are made in the discussion). For this example, the generalized inverse χ^2-method (see e.g. [15]) will be used as p value combination function:

$$\Sigma_k = \sum_{i=1}^{k} [\chi^2_{v_i}]^{-1}(1 - p_i) \tag{7}$$

where $[\chi^2_{v_i}]^{-1}$ is the inverse of a χ^2 distribution with v_i DOF, and where DOF v_i (for $i = 1, 2, ..., K$) can be chosen freely by the user. The v_i values function as weights for the stage-wise p values, with larger values corresponding to a larger weighting, i.e. when DOF v_i are increased, then the $[\chi^2_{v_i}]^{-1}(1 - p_i)$ transform will give larger values, in which case p_i will make a larger contribution towards summary statistic Σ_k (more weight is placed on stage i). It is also worth mentioning here that when $v_i = 2$ for all i, Fisher's method is obtained ($-2\ln(p_i) = [\chi^2_2]^{-1}(1-p_i)$). For the current example, v_1, v_2, and v_3 are set to 2, 3, and 4 respectively (chosen to illustrate the possibility of using distinct functions at each stage). Transforming p_i with $[\chi^2_{v_i}]^{-1}(1-p_i)$ furthermore results in a $\chi^2_{v_i}$-distributed random variable, under the condition that p_i is uniform on the [0,1] interval under H_0. For the current example, the ϕ_i distributions are therefore given by $\chi^2_{v_i}$ distributions. Next, the choice for statistical test along with the ensemble size for the first stage of the analysis needs to be chosen, after which data for stage one is collected and analysed with the statistical test of choice, thus generating p value p_1. The test can then be stopped for efficacy if $p_1 \leqslant \alpha_1$, and for futility if $p_1 \geqslant 1-\gamma_1$, else the trial proceeds to stage two of the analysis. It is worth emphasizing here that A_1 and C_1 need not be generated for the first stage of the analysis. For completeness, however, the ϕ_1 distribution (given in this example by a χ^2_2 distribution, in accordance with the choice $v_1 = 2$) is shown in Fig. 1 (plot a), along with the stage one critical boundaries A_1 and C_1. Efficacy boundary A_1 was found by solving (5), i.e. the area under ϕ_1 to the right of A_1 should equal $\alpha_1 = 0.05$, giving $A_1 = 5.992$. Futility boundary C_1 was found by solving (6), i.e. the area under ϕ_1 to the left of C_1 should equal $\gamma_1 = 0.2$, solved for $C_1 = 0.446$.

Assuming p_1 fell within the [C_1, A_1] interval, stage 2 is initiated by collecting a second group of samples. Stage two data is then analysed with the statistical test, giving p value p_2. Results from stages one and two are then combined using (7), giving $\Sigma_2 = [\chi^2_2]^{-1}(1 - p_1) + [\chi^2_3]^{-1}(1 - p_2)$, and the null distribution for Σ_2 is found using (3):

$$\phi_{\Sigma_2} = [\chi^2_2]^{T[C_1,A_1]} * \chi^2_3 \tag{8}$$

This procedure is illustrated in Fig. 1: The truncated stage one null distribution $\phi_1^{T[C_1,A_1]}$ (Fig. 1b) is convolved with ϕ_2 (Fig. 1c), giving ϕ_{Σ_2} (Fig. 1d). Note that the area under $\phi_1^{T[C_1,A_1]}$ (and consequently under ϕ_{Σ_2}) is now equal to $1 - \gamma_1 - \alpha_1 = 0.75$. Stage two critical boundaries A_2 and C_2 are again found by solving (5) and (6), respectively, i.e. the area under ϕ_{Σ_2} to the right of A_2 should equal $\alpha_2 = 0.05$, giving $A_2 = 9.695$, whereas the area under ϕ_{Σ_2} to the left of C_2 should equal $\gamma_2 = 0.4$, giving $C_2 = 4.798$. If $\Sigma_2 \leqslant C_2$ or $\Sigma_2 \geqslant A_2$, the test is stopped for futility and efficacy, respectively, else the trial proceeds to stage three.

Assuming Σ_2 fell within the $[C_2, A_2]$ interval, a third (and for this example final) group of samples is collected for the third stage of the analysis. Stage three data is then analysed, giving p value p_3, which is combined with p_1 and p_2 using (7), now giving $\Sigma_3 = [\chi^2_2]^{-1}(1-p_1)+[\chi^2_3]^{-1}(1-p_2)+[\chi^2_4]^{-1}(1-p_3)$. The null distribution for Σ_3 is then found using (4):

$$\phi_{\Sigma_3} = \phi_{\Sigma_2}^{T[C_2,A_2]} * \chi^2_4 \tag{9}$$

The procedure is again illustrated in Fig. 1: Plot (e) shows ϕ_{Σ_2} where the stage two rejection regions have been truncated, thus further reducing the area under $\phi_{\Sigma_2}^{T[C_2,A_2]}$ (and hence under ϕ_{Σ_3}) to $1-\sum_{i=1}^{2}\alpha_i+\gamma_i=0.3$, and Fig. 1f shows ϕ_3 (a χ_4^2 distribution). Convolving plots (e) and (f) gives ϕ_{Σ_3}, shown in Fig. 1g. The stage three critical boundaries A_3 and C_3 are then found using the same procedure as in stages one and two: the area under ϕ_{Σ_3} to the left of C_3 should equal to $\gamma_3=0.25$, giving $C_3=13.396$, and the area under ϕ_{Σ_3} to the right of A_3 should equal to $\alpha_3=0.05$, giving $A_3=13.396$. Note that when $\alpha+\sum_{i=1}^{K}\gamma_i=1$, that the critical boundaries for futility and efficacy at the final stage of the analysis will be the same, i.e. H_0 is either accepted for $\Sigma_3\leqslant C_3=A_3$, or rejected for $\Sigma_3\geqslant C_3=A_3$.

III. Results

This section presents results from simulations and real subject-recorded ABR data. For the simulations, the goal is to explore the performance of the CGST across a range of SNRs when using different CGST design parameters. For the subject-recorded ABR data, the goal is to provide an illustrative example of how the CGST might be used in practice.

A. Simulations

Data for the simulations consists of coloured noise with similar spectral content as real EEG background activity, along with ABR waveforms for simulating a response. The goal is to explore sensitivity and test time as a function of the SNR when using different values for K and γ_i (the stage-wise TNRs).

Method

Simulated coloured noise was generated by filtering Gaussian White Noise with an all-pole filter, where the poles of the filter were given by the parameters of an autoregressive (AR) model. The AR models were estimated from recordings of EEG background activity using the Modified Covariance method [21], with a new AR model being fit to each recording. There was approximately ∼8 hours of artefact-free EEG background activity available, which was previously recorded by [20] from 17 normal hearing adults. A total of 100 000 recordings of coloured noise were then simulated, all of which were band-pass filtered from 100-1500 Hz using a 3rd order Butterworth filter, and structured into ensembles of $N=3000$ 30.2 ms segments (henceforth referred to as epochs). Note that the 30.2 ms epochs correspond to a stimulus rate of 33.11 Hz, i.e. an additional acoustic stimulus would be presented (and an additional ABR evoked) every 30.2 ms. For each 30.2 ms epoch, an ABR was simulated by adding an appropriately rescaled ABR waveform to the epoch. The ABR waveforms were obtained from the coherent averages of subject recorded ABR data (previously collected and described in [19]), under the condition that the coherent average contained a clear response, as determined through visual inspection by an experienced audiologist (for more details, see [5]). There were a total of 34 ABR waveforms available for simulating a response. The

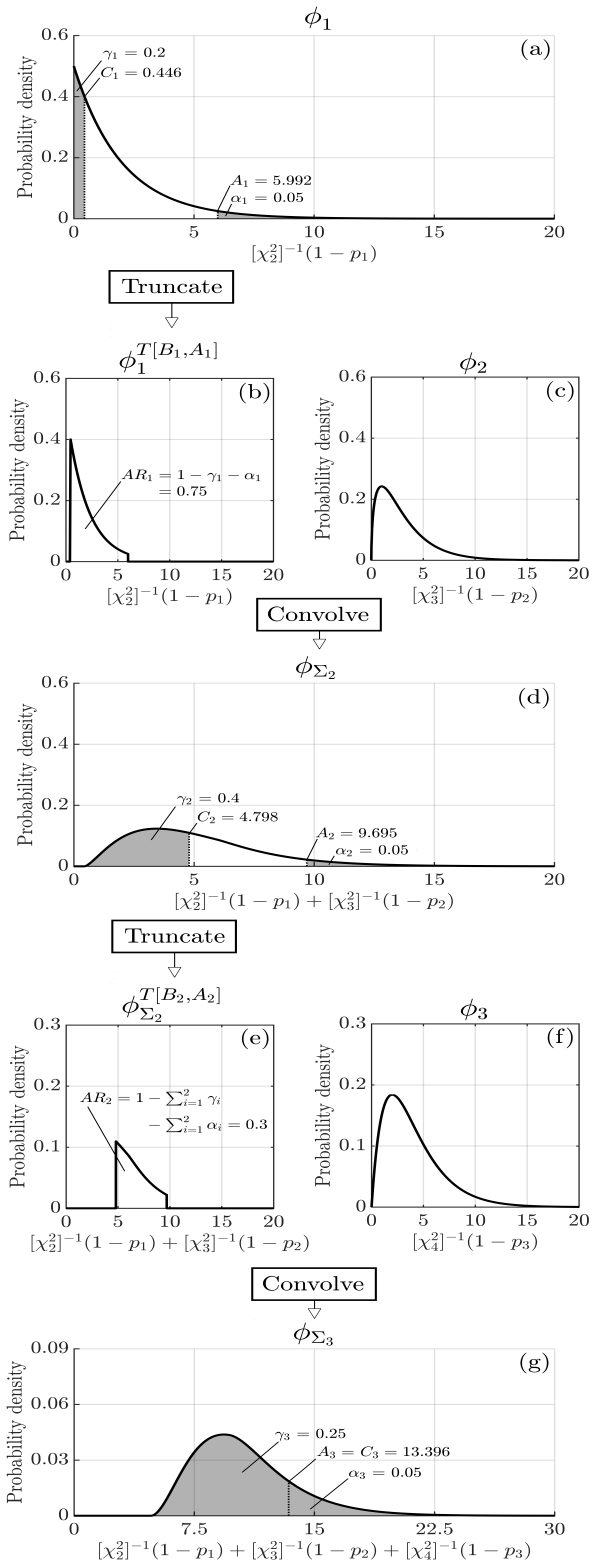


Fig. 1. An overview of the approach for generating the critical decision boundaries for a three-stage group sequential test: (a) the probability density function (pdf) for stage one, (b) the truncated pdf from stage 1, which is convolved with the pdf from stage 2 (c), giving the pdf for the stage two summary statistic (d). Truncating (d) gives (e), which is convolved with (f) to give the pdf for the stage three summary statistic in (g). Further details are presented in the text.

scaling factor for the ABR waveform was chosen such that a specific SNR was obtained, which was calculated using:

$$\text{SNR} = 10\log_{10}\frac{P_{ABR}}{P_{Noise}} \tag{10}$$

where P_{ABR} is the mean square of the (rescaled) ABR waveform, and P_{Noise} the mean square of the ensemble of epochs, prior to adding the ABR waveform and when treated as a continuous recording. The SNR was then varied from -50 dB to -20 dB, in steps of 0.5 dB. The no-stimulus condition was also included, i.e. SNR = $-\infty$.

Data were analysed in K sequential stages using the Hotelling's T^2 test (for details on analysing ABR data with the Hotelling's T^2 test, see [5]), where K took values of 1, 2, 4, or 8. The stage-wise ensemble sizes, say N_i, were all set to $\frac{3000}{K}$, i.e. the 3000 epochs were always split equally across K stages. The nominal α-level was set to 0.01, which was also split equally across the K stages, giving α_i values of $\frac{0.01}{K}$ for all i and K. Finally, the analysis was performed both with and without futility stopping. When futility stopping was used, the γ_i values were set to $\frac{0.9}{K}$, for all i and K, whereas when no futility stopping was used, the γ_i values were all set to zero.

Results

The true-positive-rates (TPRs) and mean test times (calculated across 100 000 tests) are shown in Fig. 2 as a function of the SNR, for different K, both with futility stopping (all $\gamma_i = \frac{0.9}{K}$) and without (all $\gamma_i = 0$). Results show that at high SNR (>-27.5 dB), both the single shot test and the sequential test give a 100% detection rate (both are over-powered), but that test time for the single shot test is much higher as the trial can only be stopped after the full $N = 3000$ stimuli have been presented. In particular, a reduced test time of ∼50-90% is observed (for SNRs >-27.5 dB) for the sequential test relative to the single shot test.

Results also confirm that analysing N samples using a single shot test will give a higher statistical power relative to analysing the same N samples using multiple sequentially applied statistical tests (see also [1]), i.e. a reduced TPR can be expected for increasing K for a fixed N. The reduced statistical power for the sequential test can be compensated for by increasing the ensemble size, which, in turn, increases test time. Additional simulations demonstrate that the trade-off between statistical power and test time is highly beneficial for the sequential test when detecting ABRs, i.e. for a fixed test sensitivity (achieved by varying N), the mean test time for the sequential test was reduced (relative to the single shot test) by 40-45% [6].

With respect to futility stopping, this had no noticeable effect on the TPR for these simulations (Fig. 2a and c). For relatively large SNRs (approximately >-30 dB), futility stopping also had no noticeable effect on the mean test time (Fig. 2d). For small SNRs (approximately <-30 dB), on the other hand, futility stopping resulted in very noticeable reductions in mean test time (Fig. 2d). The extent to which futility stopping affects test performance is hence dependent on the SNR of the response, but also on the choice for the γ_i values. In particular, when the evoked response has a high SNR and the γ_i values are chosen conservatively, then the Σ_i values will tend to be much larger than the C_i futility boundaries, and the test will typically not be stopped for futility. On the other hand, when the SNR is low (or a response is absent) and the γ_i values are chosen more liberally, then the Σ_i values will tend to be closer to the C_i futility boundaries, and the probability of stopping the test early in favour of H_0 is increased, potentially resulting in an increased false-negative rate (FNR). A more liberal choice for the γ_i values might therefore result in larger reductions in test time, potentially at the cost of a reduced test sensitivity.

With respect to the no-stimulus condition (SNR = $-\infty$): when no futility stopping was used, results show false-positive rates (FPRs) of 0.00949, 0.00989, and 0.00988 for $K = 2$, $K = 4$, and $K = 8$, respectively, whereas when futility stopping was used, the FPRs were 0.00953, 0.00994, and 0.00992 for $K = 2$, $K = 4$, and $K = 8$, respectively. For the single shot test ($K = 1$), a FPR of 0.0096 was observed. These results are all close to the nominal α-level of the test ($\alpha = 0.01$), and fall within the two-sided 95% confidence intervals for the expected 0.01 FPR, given by [0.0094, 0.0106]. These confidence intervals were found using a binomial distribution, constructed from 100 000 observations, where the probability of a single 'successful' Bernoulli trial (defined here as a false-positive) was set to 0.01 (the theoretical probability of a false-positive).

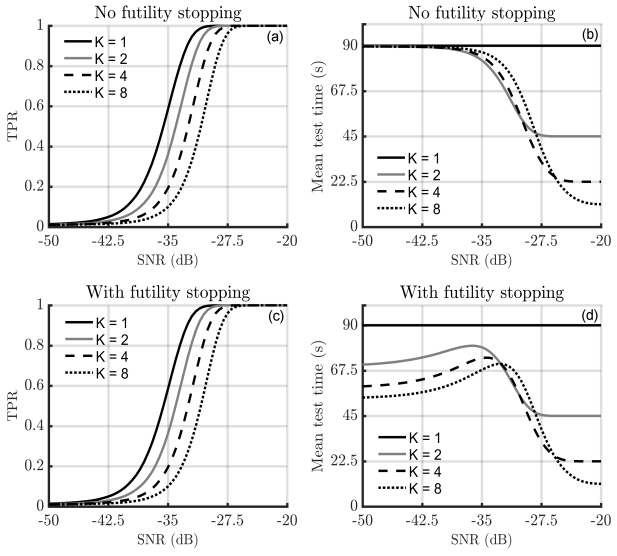


Fig. 2. Results from the simulations, which include the true-positive-rates (TPRs) and the mean test times (calculated across 100 000 tests), plotted as a function of the SNR, for various K, both with futility stopping (plots c and d) and without (plots a and b).

B. Application to Auditory Brainstem Response detection

This section provides an illustrative example of how the CGST might be used for Auditory Brainstem Response (ABR) detection. Note that the adaptive potential underlying the CGST is not explored in this section, but is considered in section B of the discussion. ABRs were previously recorded

[17] from a normal hearing adult using clicks as stimuli. The clicks were presented at a range of dB SL (sensation level) conditions, i.e. relative to the behavioural hearing thresholds. The behavioural hearing thresholds were hence first determined, achieved using a simple 'up down' approach where the amplitude of the click was decreased in steps of 10 dB for every correct response, and increased in steps of 5 dB for every missed response. The clicks were then presented to the subjects at 0, 10, 20, 30, 40, and 50 dB SL.

In total, 3000 artefact-free epochs were available for each dB SL condition. Data were then analysed using a 5-stage group sequential test. The 3000 epochs were split equally across the 5 stages, giving stage-wise sample sizes of 600 epochs. The total α-level per dB SL condition was set to 0.01, which was also split equally across the 5 stages, i.e. $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0.002$. For each dB SL condition, the fraction of tests rejected for futility was set to 0.1, 0.15, 0.2, 0.25, and 0.29 for stages 1, 2, 3, 4, and 5 respectively, i.e. $\gamma_1 = 0.1$, $\gamma_2 = 0.15$, $\gamma_3 = 0.2$, $\gamma_4 = 0.25$, and $\gamma_5 = 0.29$ (how one might choose these values is further considered in the discussion). The function for combining p values at each stage of the analysis is Fishers method [9], which can be expressed using:

$$\Sigma_k = \sum_{i=1}^{k} [\chi^2_2]^{-1}(1 - p_i) \tag{11}$$

Adopting these settings, and applying the CGST described in section two, gives the following thresholds for efficacy: $A_1 = 12.429$, $A_2 = 16.049$, $A_3 = 19.195$, $A_4 = 22.085$, $A_5 = 24.774$, along with the following thresholds for futility: $C_1 = 0.211$, $C_2 = 1.673$, $C_3 = 4.46$, $C_4 = 8.953$, $C_5 = A_5 = 24.774$. At each stage of the analysis, the subsample of 600 epochs was analysed using the Hotellings T^2 test. The resulting stage-wise p values are shown in Table 1, whereas the summary statistic Σ_i (for $i = 1, 2, ..., 5$) is shown in Table 2. The dark grey and black cells in Table 2 indicate that the trial was stopped for efficacy and futility, respectively, whereas the light grey cells indicate that the trial was allowed to proceed to the next stage of the analysis. Results from the single shot test (data were pooled across stages) confirm a detected ABR ($p < 0.01$) for the 50, 40, 30, 20, and 10 dB SL conditions, along with no detection for the 0 dB SL condition ($p = 0.6027$). The power of the sequential test is furthermore illustrated nicely in the 50 and 10 dB SL conditions, i.e. although no p value individually fell below the 0.002 threshold, the combination of multiple small p values in successive stages still resulted in H_0 being rejected. For the 0 dB SL condition, on the other hand, a succession of large p values led to an early acceptance of H_0 (additional data collection in the final stage would most likely be futile). It is also worth noting that although higher dB stimuli will tend to decrease the average detection time, it is well known that this is not guaranteed in each individual, which can be attributed to variability in the SNR, even within the same individual (due to e.g. non-stationary EEG background activity). The latter is also evident in Table 2: the 50 dB SL condition required more stages (and more stimuli) than the 20, 30 and 40 dB SL conditions before the response became statistically significant. This again emphasizes the limitations of any *a priori* choice for the sample size at any given stimulus level, even if the subjects are expected to have normal hearing. Finally, in terms of test time, early stopping at stages 4, 1, 1, 3, 4, and 4 for the 50, 40, 30, 20, 10, and 0 dB SL conditions, respectively, resulted in a total of $2400 + 600 + 600 + 1800 + 2400 + 2400 = 10200$ stimuli being used. For a stimulus rate of 33.11 Hz, this gives a total test time of $10200 \cdot \frac{1}{33.11} \approx 306$ seconds. When compared to the single shot test where the full 3000 epochs are analysed for each dB SL condition (giving a total test time of $3000 \cdot 6 \cdot \frac{1}{33.11} \approx 541$ seconds), a reduction in test time of $\sim$43% is observed.

TABLE 1
THE STAGE-WISE p VALUES GENERATED BY THE HOTELLING'S T^2 TEST, PER DB SL CONDITION, FOR A SINGLE SUBJECT.

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
50 dB SL	0.23	0.054	0.021	0.006	0.01
40 dB SL	0.001	<0.001	<0.001	<0.001	<0.001
30 dB SL	0.001	0.001	0.001	0.635	0.004
20 dB SL	0.015	0.105	<0.001	0.106	0.004
10 dB SL	0.342	0.282	0.015	0.006	0.04
0 dB SL	0.656	0.158	0.438	0.601	0.891

TABLE 2
THE SUMMARY STATISTIC Σ_i (FOR $i = 1, 2, ..., 5$) CALCULATED FROM THE STAGE-WISE p VALUES (TABLE I) USING (11), PER DB SL CONDITION, FOR A SINGLE SUBJECT. A DARK GREY CELL INDICATES THAT THE TRIAL WAS STOPPED FOR EFFICACY, A BLACK CELL INDICATES THAT THE TRIAL WAS STOPPED FOR FUTILITY, AND A LIGHT GREY CELL INDICATES THAT THE TRIAL WAS ALLOWED TO PROCEED TO THE NEXT STAGE OF THE ANALYSIS. THE DATA IN THE WHITE CELLS IS PROVIDED FOR COMPLETENESS, THOUGH THESE VALUES (AND THE DATA NEEDED TO OBTAIN THEM) WERE NOT USED WHEN REJECTING OR ACCEPTING H_0.

	Stage 1	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**
50 dB SL	2.939	8.769	16.475	26.584	35.821
40 dB SL	15.331	33.563	55.322	74.05	106.4
30 dB SL	14.866	28.341	42.726	43.634	54.634
20 dB SL	8.42	12.928	31.965	36.452	47.552
10 dB SL	2.149	4.682	13.056	23.227	29.651
0 dB SL	0.938	4.625	6.276	7.296	7.527

IV. DISCUSSION

This paper presented a novel method for finding the stage-wise critical decision boundaries (for rejecting or accepting H_0) and controlling the type-I error rate for sequentially applied statistical tests. Although originally designed for evoked response detection, the CGST can potentially be used for a wide range of applications. Indeed, the only condition for using the CGST is that the following two assumptions are satisfied: (i) the ϕ_i distributions (for $i = 1, 2, ..., K$) are mutually independent under H_0, and (ii) the ϕ_i distributions (for $i = 1, 2, ..., K$) are known *a priori*. With respect to (ii), it was assumed throughout this work that the stage-wise p values were uniform on the [0,1] interval under H_0, which is only true when the assumptions underlying the adopted statistical test are satisfied. When these assumptions are violated, then the critical decision boundaries will be inaccurate,

and increased or decreased type-I and type-II error rates can be expected. This emphasizes the importance of choosing a suitable statistical test for analysing the EEG data. For ABR detection, the reader might consider using the Hotelling's T^2 test or bootstrapped test statistics, as these have previously shown to have a good control over the FPR [5].

In general, the main advantage of using a sequential test strategy over a conventional single shot approach is a reduced mean test time, which may come at the cost of a reduced statistical power [1]. For the CGST, the trade-off between test time and statistical power is dependent on both the SNR of the response and the adopted design parameters, which include; the number of stages K; the ensemble size N; the α_i and γ_i values; and the p value transformation functions $f_i()$. Further considerations on how to choose these parameters are made in section A below.

Finally, the CGST permits a high degree of flexibility when analysing data, i.e. at each stage of the sequential analysis, all previously analysed data can be used to modify both the stage-wise ensemble sizes and the statistical test for all remaining stages. This offers new opportunities for optimising sequential test procedures in future studies. The adaptive potential underlying the CGST is further considered in section B below. Some connections between the CGST and existing methods from the literature are also drawn in section C.

A. CGST design parameters

As mentioned in section III, analysing N samples using a single test ($K = 1$) will always have a higher statistical power compared to analysing the same N samples using multiple sequentially applied tests [1]. Test time, however, will tend to be higher for the single shot test, as the test can only be stopped after the full ensemble of epochs has been collected and analysed. As shown in Fig. 2, for a fixed number of stimuli ($N = 3000$ in this case), the sequential test was highly beneficial at high SNRs (large reductions in test time, with minimal cost in test sensitivity), but less so at low SNRs (relatively small reductions in test time at the cost of large reductions in test sensitivity). If the SNR were known *a priori*, then a single shot test with an appropriate number of stimuli would be the best choice to ensure both a highly sensitive test and a short test-time. This prior knowledge is, however, typically not available, particularly so in a mixed cohort of patients tested at different stimulus intensities. Hence, in order to ensure adequate test sensitivity, one would usually err on the cautious side and include a high number of stimuli. Although this will increase the test-time for the single shot test, for the CGST such caution will have less impact, as tests with strong responses (high SNR) will still be stopped early. An additional consideration is how to split N across the K stages. Results from the current paper and [6] suggest that a relatively robust and sensitive test performance is obtained by splitting the N epochs equally across the K stages, giving stage-wise ensemble sizes of $\frac{N}{K}$.

With respect to the α_i values, there is the usual trade-off between the type-I and type-II error rate with an increase in α resulting in an increased type I and decreased type II error rate. The α_i values might therefore be chosen to optimize how statistical power accumulates throughout the sequential analysis. As an example, if the user expects a large effect size for stage one (or if N_1 is chosen to be relatively large), and a smaller effect size for stage two, they might choose to assign more α to the first stage of the analysis. If the effect size is expected to be constant throughout the test (and N is split equally across the stages), then the safe approach is to split the available α equally across the K stages, giving α_i values of $\frac{\alpha}{K}$.

With respect to the p value transformation functions f_i, these might similarly be chosen to optimise statistical power. The v_i values in the sum of inverse χ^2-distributed random variables in (7), for example, can be used as a weighting for the stage-wise p values (see section II). Again, depending on the expected effect size at each stage, v_i can be used to optimise how statistical power accumulates throughout the sequential analysis.

Finally, with respect to the γ_i values, a trade-off is introduced between statistical power and test time, i.e. larger γ_i values result in an increased probability of stopping the test early in favour of H_0, which decreases test time, potentially at the cost of an increased type-II error rate (a reduced statistical power). An additional effect associated with the γ_i values is that they reduce the remaining area under the null distribution, which affects the critical decision boundaries (for both efficacy and futility) for the remaining stages. Taking the example presented in section II, and setting $\gamma_1 = 0$ (as opposed to $\gamma_1 = 0.2$) would give stage two critical boundaries $A_2 = 9.899$ and $C_2 = 3.654$, as opposed to $A_2 = 9.694$ and $C_2 = 4.796$. Note that A_2 is reduced as γ_1 is increased, i.e. reaching statistical significance becomes easier. Hence, increasing the γ_i values can potentially reduce the risk of a type-II error. That said, the C_2 boundary for futility stopping is of course also increased, which increases the probability of a false-negative. In general, increasing the γ_i values will indeed tend to result in a reduced statistical power and a reduced test time. As shown in the simulations (section III), a relatively safe choice for the γ_i values (i.e. minimal loss in test sensitivity) is to split the available γ equally across the K stages, giving γ_i values of $\frac{1-\alpha}{K}$.

B. An adaptive group sequential test

An adaptive group sequential test is a repeated testing procedure (applied to sequentially acquired groups of samples) that allows test parameters to be modified throughout a trial without compromising the overall type-I error rate [28]. Examples of the type of adaptations permitted include the stage-wise sample sizes (see e.g. [17, 23]), modifications to the number of remaining tests within the trial (e.g. [15]), and potentially even a change in the choice of statistical test.

Various adaptive group sequential tests can be found in the literature, the majority of which are built around either (1) conditional error functions [16,22,23], i.e. the conditional probability of incorrectly rejecting the null hypothesis given the test statistic from the previous stage, or (2) analysing the data in disjoint sub-samples and finding an appropriate

critical decision boundary for some combination function of the stage-wise p values [1, 2, 4, 15, 26]. These methods differ primarily in terms of complexity and flexibility. The earlier designs in [1] and [23], for example, are limited in regards to both the number of stages permitted and the choice for the stage-wise critical decision boundaries. The methods following these earlier designs strive to either simplify the construction of adaptive group sequential tests [25], or to provide additional design flexibility in terms of the choice for critical decision boundaries and the type of adaptations permitted [2,4,15-17,22].

With respect to the CGST, all previously analysed data can be used to choose the sample size and the statistical test (including the statistical features) for the remaining stages of the sequential analysis. The stage-wise critical decision boundaries A_i and C_i (for $i = 1, 2, ..., K$), however, should be designed *a priori*, i.e. independently of the data being analysed. This implies that the following CGST design parameters should be fixed in advance: the total number of stages to perform K; the stage-wise type-I error rates α_i; the γ_i values for futility; and the p value transformation functions $f_i()$. Note again that the CGST only uses the p values generated by the statistical analyses. The A_i and C_i boundaries are therefore not dependent on N, neither are they dependent on choice for statistical test. Consequently, both the sample size and the choice for statistical test can be adapted throughout the sequential analysis without introducing a bias.

C. Some connections to existing methods

For auditory evoked response detection, an alternative sequential test strategy has previously been proposed by Stürzebecher et al (2005) in [26]. In this approach, data is continuously being pooled in a single ensemble, which is re-analysed at various pre-determined time intervals. The critical decision boundaries are then found *a priori* using Monte-Carlo simulations. An important difference between this approach and the CGST is the independence assumption between each stage of the sequential analysis, which is not required in Stürzebecher et al (2005). As a result, the approach in Stürzebecher et al (2005) does not permit data-driven adaptations.

Various connections between the CGST and some adaptive methods from the literature firstly include the class of 'self designing tests' described by Hartung & Knapp (2003) in [15]. In Hartung & Knapp, data is analysed in disjoint groups of samples (as is the case with the CGST), and a p value is generated at each stage of the sequential analysis. The stage-wise p values are then combined using the generalized inverse χ^2-method (see also (7) in section II), and the null hypothesis H_0 can be rejected at stage k when summary statistic Σ_k exceeds some threshold A_{v_Σ}, i.e. $\Sigma_k > A_{v_\Sigma}$. Note that, unlike the CGST, there is now just a single critical decision boundary, which is calculated directly using:

$$A_{v_\Sigma} = [\chi^2_{v_\Sigma}]^{-1}(1-\alpha) \tag{12}$$

where v_Σ are the DOF of a χ^2 distribution, chosen freely (prior to the test) by the user. The DOF v_Σ essentially functions as a 'currency' that the user is free to 'spend' throughout the trial, e.g. at stage i, the user should specify degrees of freedom v_{i+1}, which is then used to transform p_{i+1} into a $\chi^2_{v_{i+1}}$-distributed random variable, after which it is combined with all previously generated (and χ^2-transformed) p values (see (7)). The main advantage for this approach over some alternatives is that the number of stages K need not be specified. Instead, the user is free to spend v_Σ until it has been depleted, i.e. until $\sum_{i=1}^{K} v_i = v_\Sigma$. A potential disadvantage for this approach is that early stopping in favour of H_0 is not permitted. Note also that the stage-wise type-I error rates are 'hidden' from the user, i.e. how statistical power accumulates throughout the trial is not transparent. When using the CGST, on the other hand, the user is given the choice to explicitly specify the stage-wise type-I error rates (through the α_i values), which makes the choice easier to understand and hence optimize.

Connections with additional adaptive methods worth mentioning include the 'sum of p values' approach described by Chang (2007) in [4], which can be represented by the CGST by setting the combination function to $\Sigma_k = \sum_{i=1}^{k} w_i p_i$, where w_i is the chosen weight for stage i. The ϕ_i distributions are then uniformly distributed on $[0, w_i]$ under H_0. The CGST also represents the class of adaptive group sequential tests described by Bauer & Köhne in [1], achieved by using Fishers method as p value combination function, and by choosing appropriate values for K, α_i, and γ_i.

V. Conclusion

The CGST is a flexible and intuitive method for finding the stage-wise critical decision boundaries and controlling the type-I error rate of sequentially applied statistical tests. Although originally designed for evoked response detection, the CGST can be used for a wide range of sequential test applications, albeit under the condition that the stage-wise p value null distributions (the ϕ_i distributions) are both mutually independent and known *a priori*. In general, sequential testing introduces trade-offs between statistical power and test time. For the CGST, this trade-off is dependent on both the SNR of the response and the choice of CGST design parameters. A suitable selection of CGST design parameters is therefore essential when optimising test performance. The CGST furthermore falls under the class of 'adaptive group sequential tests'; a category of sequential test strategies that permit data-driven adaptations to test parameters throughout the sequential analysis. For the CGST, adaptations to the sample size and the statistical test are permitted, which can be explored in future studies when further optimising sequential test procedures. Finally, as shown in the discussion, the CGST is a generalized form of some alternative adaptive group sequential tests found in the literature, and one that facilitates understanding and permits a high flexibility in the choice of strategy.

Acknowledgment

This work was supported by the Oticon Foundation and the Engineering and Physical Sciences Research Council (EPSRC, grant No. EP/M026728/1). The authors would also like

to acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work. The subject recorded ABR data used in section 3 is openly available at tat the University of Southampton repository at http://doi.org/10.5258/SOTON/D0168.

REFERENCES

[1] P. Bauer, and K. Köhne, Evaluation of experiments with adaptive interim analyses. *Biometrics*, vol. 50(4), pp. 1029-1041, 1994. DOI: 10.2307/2533441

[2] W. Brannath, M. Posch, and P. Bauer, Recursive combination tests. *J. Am. Stat. Assoc.*, vol. 97 (457), pp. 236-244, 2002. DOI: 10.1198/016214502753479374

[3] M. Cebulla, E. Stüurzebecher, and C. Elberling, Objective detection of Auditory Steady State Responses: Comparison of One-Sample and q-Sample Tests. *J. Am. Acad. Audiol*., 17(2), pp. 93-103, 2006. DOI: 10.3766/jaaa.17.2.3

[4] M. Chang, Adaptive design method based on sum of p-values, *Statist. Med*., vol. 26(14), pp. 2772-2784, 2006. DOI: 10.1002/sim.2755

[5] M. A. Chesnaye, S. L. Bell, J. M. Harte, and D. M. Simpson, Objective measures for detecting the auditory brainstem response: comparisons,of specificity, sensitivity and detection time. *International Journal of Audiology*, vol. 57(6), pp. 468-478, 2018. DOI: 10.1080/14992027.2018.1447697

[6] M. A. Chesnaye, S. L. Bell, J. M. Harte, and D. M. Simpson, A group sequential test for auditory brainstem response detection. *International Journal of Audiology*, *under revision.*

[7] S-C. Chow, and M. Chang M, Adaptive Design Methods in Clinical Trials. Chapman & Hall/CRC Biostatistics Series. p.156, 2007.

[8] J. Cohen, Statistical Power Analysis for the Behavioural Sciences. 2nd ed. Hillsdale, N.J. : L. Erlbaum Associates, 1988.

[9] E. Colon, and S.L. Visser, *Evoked Potential Manual. A Practical Guide to Clinical Applications.* Springer, The Netherlands, 1990. DOI: 10.1007/978-94-009-2059-0

[10] C. Elberling, and Don M. Quality estimation of averaged auditory brainstem responses. *Scandinavian audiology*, 13(3), pp. 187-197, 1984. DOI: 10.3109/14992028409043059

[11] R. A. Fisher, Statistical methods for research workers, 11th ed. Oliver and Boyd, Edinburgh, 1932.

[12] M. Golding, H. Dilon, J. Seymour, and L. Carter, The detection of adult cortical auditory evoked potentials (CAEPs) using an automated statistic and visual detection. *International Journal of Audiology*, 48(12), pp. 833-842, 2009. DOI: 10.3109/14992020903140928.

[13] C. M. Grinstead, and J. L. Snell, Chapter 7, in Introduction to Probability, 2nd ed., American Mathematical Society, the United States of America, pp. 285, 1997.

[14] J.W. Hall, *New Handbook of Auditory Evoked Responses*. 1st ed. London: Pearson. p.95, 2006

[15] J. Hartung, and G. Knapp, A new class of completely self-designing clinical trials. *Biometrical J*., vol 45(1), pp. 3-19, 2003. DOI: 1002/bimj.200290014

[16] Q. Liu, and G. Y. H. Chi, On sample size and inference for two-stage adaptive designs, *Biometrics*, vol. 57(1), pp. 172-177, 2001. DOI: 10.1111/j.0006-341X.2001.00172.x

[17] W. Lehmacher, and G. Wassmer, Adaptive sample size calculations in group sequential trials, *Biometrics*, vol. 55(4), pp. 1286-1290, 1999. DOI: 10.1111/j.0006-341X.1999.01286.x

[18] R. C. Littell, and J. L. Folks, Asymptotic Optimality of Fishers Method of Combining Independent Tests, *J. Am. Stat. Assoc*., vol. 66(336), pp. 802-806, 1971.

[19] J. Lv, D. M. Simpson, and S. L. Bell, Objective detection of evoked potentials using a bootstrap technique, *Med. Eng. Phys.*, vol. 29(2), pp. 191-198, 2007.

[20] Madsen S.M.K., Harte J.M., Elberling C. & Dau T. (2017). Accuracy of averaged auditory evoked potential amplitude and latency estimates. *International Journal of Audiology*, 57(2), pp. 1-9.

[21] Marple S.L.Jr. Digital Spectral Analysis with Applications. Prentice-Hall, Englewood Cliffs, NJ, 1987

[22] H.H. Müller, and H. Shäfer, Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical,Group Sequential Approaches, *Biometrics*., vol. 57(3), pp. 886-891, 2001. DOI: 10.1111/j.0006-341X.2001.00886.x

[23] M. A. Proschan, and S. A. Hunsberger, Designed extension of studies based on conditional power, *Biometrics*, vol. 51(4), pp. 1315-1324, 1995. DOI: 10.1016/0197-2456(95)91243-6

[24] D.M. Simpson, C.J. Tierra-Criollo, R.T. Leite, E.J.B. Zayen, & A.F.C Infantosi, Objective Response Detection in an Electroencephalogram During Somatosensory Stimulation. *Annals of Biomedical Engineering*, Vol 28(6), pp. 691698, 2000. DOI: 10.1114/1.1305530.

[25] J. Sheng, and L. Qiu, p-Value calculation for multi-stage additive tests, *J. Stat. Comput. Sim*., vol. 77(12), pp. 10571064, 2007. DOI: 10.1080/10629360600872707

[26] E. Stürzebecher, M. Cebulla, & C. Elberling., Automated auditory response detection: Statistical problems with repeated testing. *International Journal of Audiology*, 44(2), pp. 110-117, 2005.

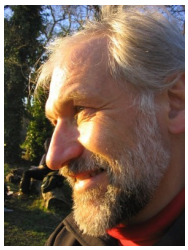
[27] Walsh P., Kane N. & Butler S, The clinical role of evoked potentials. *Journal of Neurology, Neurosurgery & Psychiatry*, Vol 76(suppl 2), pp. 16-22, 2005. DOI: 10.1136/jnnp.2005.068130

[28] G. Wassmer, Basic concepts of group sequential and adaptive group sequential test procedures, *Statistical papers*, vol. 41(3), pp. 253-279, 2000. DOI: 10.1007/BF02925923

Michael A. Chesnaye received his PhD degree in Biomedical Engineering at the Institute for Sounds and Vibration Research at the University of Southampton in 2019. His main research interests are related to signal processing, statistics and pattern recognition of neurophysiological signals.

Steven L. Bell is an Associate Professor of Audiology in the Hearing and Balance Centre, Institute of Sound and Vibration Research, University of Southampton and is a registered Clinical Scientist (audiology). His main areas of research interest are auditory and vestibular evoked potentials. This includes improving detection of responses, clinical applications, and in furthering understanding of hearing and balance function. He is also interested in objective methods to evaluate the performance of assistive devices, such as hearing aids and cochlear implants.

James M. Harte is the Director of the Interacoustics Research Unit, located at the Technical University of Denmark. He has held various academic positions in the University of Warwick, Technical University of Denmark and the University of Southampton. His research interests are in biomedical signal processing and technical audiology, aiming to turn basic understanding of the mechanical transduction and neural coding mechanisms of the auditory system into novel methods for diagnosing hearing impairment.



David M. Simpson After schooling in Austria, David Simpson graduated in Biomedical Electronics from the University of Salford (1981), and then worked as a mathematics and physics teacher in Nigeria. He obtained his PhD in Electrical Engineering from Imperial College of Science, Technology and Medicine, University of London in 1988, initiating his research interests in signal and image processing with applications in medicine. From 1989 to 1998 he was a lecturer in the Biomedical Engineering Program at the Federal University of Rio de Janeiro (Brazil). He then became a research fellow in the Medical Physics Department of Leicester Royal Infirmary before moving to the University of Southampton (ISVR) in 2001, where he is Professor of Biomedical Signal Processing and Head of the Human Sciences Group in the Faculty of Engineering and Physical Sciences. Research interests are focused on signal processing in neurophysiology, as well as cerebro- and cardiovascular control.