Biomedical Signal Processing and Control 108 (2025) 107937

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Bayes factors for sequential auditory brainstem response detection

M.A. Chesnaye [a,b,*], D.M. Simpson [b], J. Schlittenlacher [c], S. Laugesen [d], S.L. Bell [b]

[a] National Acoustic Laboratories, Hearing Australia, Sydney, Australia
[b] Faculty of Engineering and the Environment, Institute of Sound and Vibration Research, University of Southampton, Southampton, United Kingdom
[c] Division of Psychology and Language Sciences, University College London, London, United Kingdom
[d] Interacoustics Research Unit, C/O Technical University of Denmark, Lyngby, Denmark

ARTICLE INFO

Keywords:
Auditory Evoked Potentials
Auditory Brainstem Response
Objective Detection Methods
Sequential Testing
Bayes Factors

ABSTRACT

Objective: When determining the presence or absence of an Auditory Brainstem Response (ABR), clinicians often visually inspect the accruing data over time, i.e., a sequential test is adopted. The current work presents and evaluates Bayes Factors (BFs) as an objective sequential test for assisting clinicians with this task.
Method: Test specificity and sensitivity were optimised in simulated data and evaluated in subject-recorded data, including no-stimulus recordings (17 adults) and chirp-evoked ABR recordings (31 adults, 9 with hearing loss). The BF approach was compared with an existing sequential test, called the Convolutional Group Sequential Test (CGST).
Results: In simulations, BFs reduced mean test times by 60–70 % relative to the CGST while maintaining equal sensitivity and specificity. Similar reductions were observed in subject-recorded EEG background activity (~70 %) and in chirp-evoked ABRs (0–60 %, depending on stimulus levels). For BFs, test time is tied to noise levels in the data, which allows test sensitivity to be controlled even when noise levels are high. The drawback is that the FPR is also tied to test time, and results show small variations (<0.01) in FPRs depending on noise levels. In contrast, test time for the CGST is fixed, giving an improved control over the FPR, but a reduced control over test sensitivity.
Significance: BFs demonstrated high sensitivity and reduced mean test times relative to the CGST. It also provides regular feedback with no maximum test time specified, making it well-suited at assisting clinicians with different levels of expertise and feedback preferences.

1. Introduction

The Auditory Brainstem Response (ABR) is a brief change in neural activity generated along the auditory pathway in response to sound [1]. It can be measured non-invasively using scalp electrodes and is routinely used in hearing screening and audiogram estimation in newborns and other hard-to-test populations[2,3] as well as detecting some neurological disorders (e.g., [4]). Usually, the first step in these applications is to determine whether an ABR is present or absent, after which additional analysis can be carried out on the amplitude and morphology of the ABR.

One challenge with detecting the ABR is that it is often hidden in the background activity, which can be an order of magnitude larger than the ABR [1]. To reliably detect the ABR, it is therefore important to first improve its Signal-to-Noise Ratio (SNR), which is achieved by presenting many stimuli to the subject and averaging the brief time-intervals following stimuli onsets to reduce "noise". An experienced clinician then visually inspects the averaged waveforms to determine if a response is present or absent [2]. A drawback with visual inspection, however, is that results can vary both within and between examiners [5,6], thus introducing a subjective, examiner-dependent element to the analysis, potentially compromising the accuracy of the test.

In most ABR applications, errors in decision-making can have significant implications: An ABR that is incorrectly deemed present, for example, may lead to undiagnosed hearing loss. To reduce subjectivity and increase test accuracy and efficiency, researchers have sought to automate the procedure through objective detection methods (e.g. [7,8,9,10,11,12,13]). In the literature, these methods are often evaluated as "single shot" tests, which implies that data is analysed just once. This contrasts with applications in the clinic where methods are applied repeatedly to the accruing data over time, known as a sequential test. Sequential tests are important for providing timely feedback to

* Corresponding author.
E-mail address: michael.chesnaye@nal.gov.au (M.A. Chesnaye).

https://doi.org/10.1016/j.bspc.2025.107937
Received 10 August 2024; Received in revised form 28 March 2025; Accepted 14 April 2025
Available online 15 May 2025

examiners. They also help to keep test time low as data collection can be stopped early for the high SNR responses. When compared to single shot test strategies, sequential tests previously reduced mean test times for ABR detection by up to approximately 45 % whilst maintaining equal test sensitivity [11,12].

Sequential tests have clear advantages over single shot test strategies but present a challenge in terms of controlling the False-Positive Rate (FPR). This is because repeated hypothesis testing increases the chance of finding spurious effects in noise, which means that the False-Positive Rate (FPR) increases as more tests are carried out, known as an inflated FPR [14]. To prevent inflated FPRs, the critical thresholds for response detection need to be chosen carefully, for which various methods have been proposed (e.g., [9,12,15,16,17,18]). However, many of these methods require the statistical analysis to be pre-specified in terms of how often and when data is analysed, leading to relatively inflexible test protocols that may not be optimally efficient. Note also that if the sequential test comes to an end without having reached a clear outcome in terms of ABR present or absent, then the test cannot be prolonged, at least not without also inflating the FPR.

The aim for the current study was to introduce and evaluate a flexible and efficient sequential test for assisting clinicians with ABR detection, built around the Bayes Factor (BF). The BF is a measure of the evidence for one hypothesis over another [19,20], defined in the current study as "H_0: ABR is absent" and "H_1: ABR is present". More specifically, the BF is a ratio of likelihoods, defined as the likelihood that the data arose under H_1 over the likelihood that the data arose under H_0. If H_1 is true, then the BF value increases towards infinity as data accrues, whereas if H_0 is true, then the BF value decreases towards zero. Statistical inference is then carried out by placing upper and lower thresholds on the BF, i.e., H_1 is accepted for $BF > BF_{High}$, or H_0 for $BF < BF_{Low}$, else (for $BF_{Low} > BF < BF_{High}$) further data collection is deemed necessary before making a final decision.

The BF approach is attractive for ABR detection, firstly because the probability of inferring a false-positive decreases as data accumulates, which implies that the FPR cannot exceed some fixed upper threshold, regardless of how often or how long data is analysed. This contrasts with conventional frequentist analyses where controlling the FPR requires limitations to be imposed on the frequency and duration of the analysis. The practical implication is that the BF approach can be used to assist with data analysis for as long as the clinician deems necessary. The BF approach can also be applied to the accruing data frequently (e.g, every 3 seconds), providing regular feedback. Lastly, the BF approach aligns with the intuitions of the examiners who are similarly expected to make fewer errors as data accrues, potentially resulting in a more trustworthy detector.

The BF approach was previously also applied successfully in the related field of Auditory Steady State Response (ASSR) detection [21], where it was compared to the Neyman-Pearson (NP) detector and a modified Sequential Probability Ratio Test. All three methods use a likelihood ratio, computed between two competing hypotheses, and differences between methods lie primarily in how the critical thresholds are constructed. The BF approach in the current work is similar to methods in [21] but differs in that it uses two thresholds – one for accepting H_0 (response absent) and one for accepting H_1 (response present). This contrasts with [21] where thresholds were defined for accepting H_1 only, i.e. early stopping in favour of H_0 was not considered. It is also worth noting that methods in [21] were evaluated as single shot tests, rather than as sequential tests. The current study thus extends work from [21] to ABR detection whilst also introducing inference on both H_0 and H_1. The procedure is then evaluated within a sequential testing framework.

In the following sections, the BF approach is described in more detail and then evaluated in terms of specificity, sensitivity, and efficiency (the required sample size for response detection). Data for the assessment comprised realistic simulated data, as well as recordings of EEG background activity from 17 adults with normal hearing, and chirp-evoked ABR recordings from 31 adults, 9 of which had some degree of hearing loss. Comparisons were also drawn with an existing sequential test strategy from [11] – the Convolutional Group Sequential Test (CGST) – which pre-specifies when and how often to analyse data. Some pros and cons underlying the BF approach are considered in the Discussion along with directions for future work.

2. Methods

ABR measurements involve presenting many stimuli to a subject and recording EEG in the short time intervals following stimulus onset. These short time intervals are referred to as "epochs" and consist in the current work of voltage measurements along the 0–15 ms post-stimulus intervals. In matrix format, epochs are represented as:

$$\mathbf{D} = \begin{bmatrix} d_{11} & \cdots & d_{1J} \\ \vdots & \ddots & \vdots \\ d_{N1} & \cdots & d_{NJ} \end{bmatrix} \tag{1}$$

where d_{ij} is the jth sample of the ith epoch, N is the total number of epochs (the ensemble size), and J is the number of samples within each epoch. The mean epoch, known as the coherent average, is found by averaging down each of the J columns. This is typically inspected visually by examiners to determine whether it contains an ABR or not. Additionally, data can be analysed using some statistical detection method.

2.1. Bayes Factors for ABR detection

The Bayes Factor (BF) is a measure of the strength of evidence for one hypothesis compared to another [19]. In the current work, these hypotheses are the null hypothesis, "H_0: ABR is absent", and the alternative hypothesis, "H_1: ABR is present". The evidence is an F statistic, extracted from data matrix **D** using the time domain Hotelling's T^2 (HT^2) test [10,22]; see also Section 2.2.1). The BF value itself can be expressed as:

$$BF = \frac{L(F|H_1)}{L(F|H_0)} \tag{2}$$

where $L(F|H_0)$ and $L(F|H_1)$ are the likelihoods that F arose if H_0 or H_1 were true, respectively. Statistical inference can then be carried out by constructing upper and lower thresholds for the BF value: If $BF > BF_{High}$, then H_1 is accepted as true, whereas if $BF < BF_{Low}$, then H_0 is accepted, else the result is inconclusive and additional data collection is deemed necessary. An illustrative example of how the BF value is computed is also shown in Fig. 1, panel (a).

The main components of the BF approach thus comprise: (1) the test statistic, given in the current work by an F statistic, (2) the distributions of the F statistic under the H_0 and H_1 hypotheses, which are needed to compute the $L(F|H_0)$ and $L(F|H_1)$ components, and (3) the BF_{High} and BF_{Low} critical thresholds for inference. Note also that in order to define the H_1 hypothesis, assumptions need to be made about the ABR. These components are further considered in the sections below.

2.1.1. The F statistic

The F statistic was extracted from **D** using the time domain HT^2 test. First, each 15 ms epoch in **D** was compressed into Q = 25 "voltage-means", which involved splitting each epoch into 25 non-overlapping 0.6 ms segments, and averaging across segments to give an Nx25-dimensional matrix of voltage means:

$$\mathbf{V} = \begin{bmatrix} v_{11} & \cdots & v_{1Q} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NQ} \end{bmatrix} \tag{3}$$

where v_{ij} is the jth voltage mean extracted from the ith epoch. A T^2 statistic is then computed using [23]:

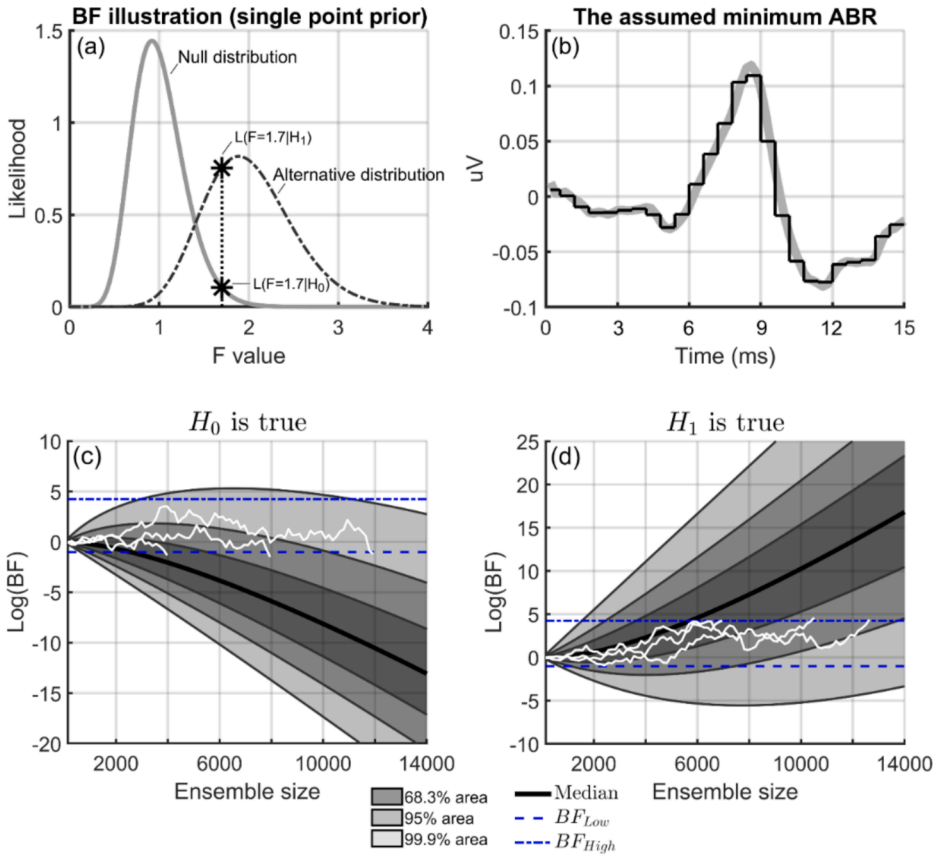

Fig. 1. Panel (a) illustrates how the Bayes Factor (BF) is computed when using single point priors for the null and alternative hypotheses, denoted by H_0 and H_1, respectively. In this example, Q = 25 voltage means were extracted from an ensemble of N = 1000 epochs, which were then analysed by the Hotelling's T^2 test, giving an F value of 1.7. Under H_0, F was assumed to follow a central F distribution with Q and N degrees of freedom, whereas under H_1, F was assumed to follow a non-central F distribution Q and N degrees of freedom along with non-centrality parameter $\lambda = 0.025$. The λ parameter was extracted from an ABR template waveform (Section 2.1.4), shown in **panel (b)**. The Q = 25 voltage means ($\boldsymbol{\mu}$ in Eq. (7)) extracted from this waveform were overlayed for illustration purposes. Using Eq. (2), the BF value is then given by $BF = \frac{L(F|H_1)}{L(F|H_0)} = \frac{0.754}{0.104} = 7.25$, i.e., the numerator, $L(F = 1.7|H_1)$, is given by the height of the alternative distribution at location F = 1.7 and equals 0.754, whereas the denominator, $L(F = 1.7|H_0)$, is given by the height of the null distribution at location F = 1.7 and equals 0.104. **Panels (c)** and **(d)** illustrate the expected trajectories for the log-transformed BF values under H_0 and H_1, shown as shaded regions. For illustration purposes, three hand-selected example trajectories under each hypothesis were also shown as white lines, along with the log-transformed BF_{High} threshold for accepting H_1 (here equal to 4.23) and the log-transformed BF_{Low} threshold for accepting H_0 (equal to −1.01). Testing continues until the BF value exceeds either BF_{High} or BF_{Low}, in which case an ABR would be deemed present or absent, respectively.

$$T^2 = N \bullet \mathbf{x} \bullet \mathbf{S}^{-1} \bullet \mathbf{x} \tag{4}$$

where $\mathbf{x}$ is a Q-dimensional vector of "mean voltage-means", found by averaging down the Q columns of $\mathbf{V}$, $\mathbf{S}^{-1}$ is the inverse of the covariance matrix of $\mathbf{V}$, and' denotes vector transpose. Finally, the T^2 statistic is transformed into an F statistic [23]:

$$F = T^2 \frac{N - Q}{Q(N - 1)} \tag{5}$$

which follows a central F distribution under H_0, and a non-central F distribution under H_1.

It is worth noting that the T^2 statistic in Eq. (4) is a measure of the signal amplitude, represented by $\mathbf{x}$, relative to the noise, represented by covariance matrix $\mathbf{S}$, and is thus related to the SNR. It is, however, a special type of SNR because, unlike standard SNR calculations that consider signal and noise power, the T^2 statistic also incorporates the covariance of the noise. In situations where features are correlated, this results in a more powerful test compared to tests that consider just the variance. For more in-depth discussion of the HT^2 test, see also [24] and Rencher et al. (2001).

The number of voltage-means, Q, to extract from the 15 ms analysis window is an important parameter that impacts on test performance. When Q is too small, peaks and troughs in the ABR waveform may cancel out due to averaging, leading to reduced signal amplitudes and potentially a less powerful test. While this can be prevented by increasing Q, a larger Q comes with the adverse effect of a higher dimension of the feature matrix $\mathbf{V}$. This is undesirable, as it introduces many unknown parameters (variances and covariances) to the analysis, all of which need to be estimated from the data, potentially resulting in additional uncertainty and a less powerful test (e.g. [25]). The choice for Q = 25 in the current work was motivated by findings in [10] where Q = 25 led to a good test sensitivity when combined with a 15 ms analysis window. Finally, note that the averaging procedure functions as a low-pass filter with down sampling. When averaging across 0.6 ms segments, the effective sampling rate is 1000/0.6 = 1.67 kHz (rather than the original 5 kHz) with a Nyquist frequency of 835 Hz, thus limiting the bandwidth of the voltage means to 30–835 Hz, rather than the 30–1500 Hz band of the original data.

2.1.2. The null distribution

If H_0 is true (no ABR is present), then the expected values for $\mathbf{x}$ are given by a Q-dimensional vector of zeros because (1) the ABR amplitude is assumed to be zero, and (2) the direct current component (mean signal value) is removed from the data during the data pre-processing stage by a high-pass filter (a 3rd-order Butterworth filter with a 30 Hz cut-off

frequency in the current work). In this case, F is assumed to follow a central F distribution with Q and N-Q degrees of freedom [23]. The $L(F|H_0)$ term in the denominator of Eq. (2) for computing the BF is accordingly given by:

$$L(F|H_0) = \varnothing_C(F, Q, N-Q) \tag{6}$$

where $\varnothing_C$ denotes a central F distribution. Note that Eq. (6) describes the height of $\varnothing_C$ at location F (panel a, Fig. 1).

2.1.3. The alternative distributions

If H_1 is true (an ABR is present), then the F statistic is assumed to follow a non-central F distribution with Q and N-Q degrees of freedom along with non-centrality parameter λ [23]:

$$\lambda = N \bullet \boldsymbol{\mu} \bullet \Sigma^{-1} \bullet \boldsymbol{\mu} \tag{7}$$

where **μ** is a Q-dimensional vector containing the expected values for **x** (representing the ABR), and Σ is a QxQ dimensional matrix containing the expected values for **S** (representing the background activity). The λ parameter is thus equivalent to the T^2 statistic in Eq. (4), after substituting **x** and **S** for their expected values, **μ** and Σ, and is therefore also closely related to the ABR's SNR.

In the ideal scenario, both **μ** and Σ would be known before the test, allowing H_1 to be defined using a single, subject- and recording-specific λ value. The challenge, however, is that both **μ** and Σ are typically not known, and to define the alternative distribution, assumptions must be made. One approach is to assume λ directly, and to define a prior distribution across the range of λ values typically observed under H_1. The drawback, however, is that H_1 is then no longer explicitly tied to the ABR, i.e., because λ is an SNR-based metric, each value may correspond to a multitude of ABR and noise power combinations. As a result, statistical inference no longer directly informs the clinician about the actual amplitude of the ABR, only its amplitude relative to the noise.

Instead of specifying λ directly, it was opted to estimate Σ using **S**, and to extract the **μ** values from an ABR template waveform, rescaled to achieve a certain peak to trough amplitude (PTTa) value. Doing so allows the λ parameter (and hence the H_1 hypothesis) to be defined as a function of the ABR's PTTa. Using the ABR PTTa is attractive as it is a commonly used metric in audiology. For example, clinicians apply the PTTa-based "3 to 1 rule" when visually detecting ABRs [2], requiring the PTTa value to be at least three times the estimated noise before a response is deemed present. There is also substantial literature on ABR PTTa values (e.g., [26]), which facilitates the construction of priors over PTTa values (considered below) and may ultimately lead to an intuitive and interpretable test that is more readily integrated with current clinical practices.

Regarding the template waveform, this was constructed by averaging ABR measurements from 12 adults with normal hearing [27] – a more detailed overview of the averaging process is given in the **Supplemental Digital Content**. The template waveform was then rescaled to achieve a certain PTTa value, say A. Note that this PTTa value corresponds to the difference between the largest peak and the smallest trough of the ABR template, which is dominated by wave V. Prior distributions were then placed over A, covering PTTa values ranging from 0.2, up to 1.6 μV, i.e. the range of ABR amplitude values typically observed in the literature (e.g., [26]). To also assess the sensitivity of the approach to the choice of prior, a total of four prior distributions were considered.

1. *Single point prior*

In the most straightforward case, a "single point prior" can be used, which considers a single "minimum ABR waveform" with a PTTa value of 0.2 μV. This prior, shown in panel (a) of Fig. 2, can be modelled using a Dirac delta function, which is infinite at 0.2 μV, and zero everywhere else:

$$Pr(A|H_1) = \delta(A - 0.2) \tag{8}$$

where δ is the Dirac delta function. Interpreting the output of the BF approach under the single point prior is relatively intuitive: If $BF < BF_{Low}$, then H_0 is accepted, and it is concluded that an ABR was absent, or that the ABR PTTa value was smaller than the minimum amplitude of 0.2 uV. If $BF > BF_{High}$, then it is concluded that an ABR is present and that the PTTa value of the ABR was at least 0.2 uV.

2. *Truncated exponential prior*

A potential drawback for the single point prior is that the $L(F|H_1)$ term (the numerator in Eq. (2) may yield relatively low likelihood values in cases where the ABR PTTa values exceed 0.2 uV, potentially leading to reduced test sensitivities. To address this, a distribution can be placed over a range of hypothesised PTTa values. The first distribution that was considered was a truncated exponential distribution, which was defined so that 99 % of its area lay along the 0.2 to 1.6 μV interval:

$$Pr(A|H_1) = \frac{1}{0.304}\exp\left(-\frac{A-0.2}{0.304}\right) \tag{9}$$

where all probability densities were set to zero for $A < 0.2$, i.e., the distribution was truncated to the [0.2, 1.6] uV interval – see Fig. 2, panel (b). Under this distribution, ABR PTTa values of 0.2 μV were deemed most probable, with probability densities decreasing exponentially for increasing PTTa values.

When placing a distribution over A, note that multiple ABR templates are now considered when computing $L(F|H_1)$ in the BF equation. In particular, computing $L(F|H_1)$ requires integrating over the PTTa A:

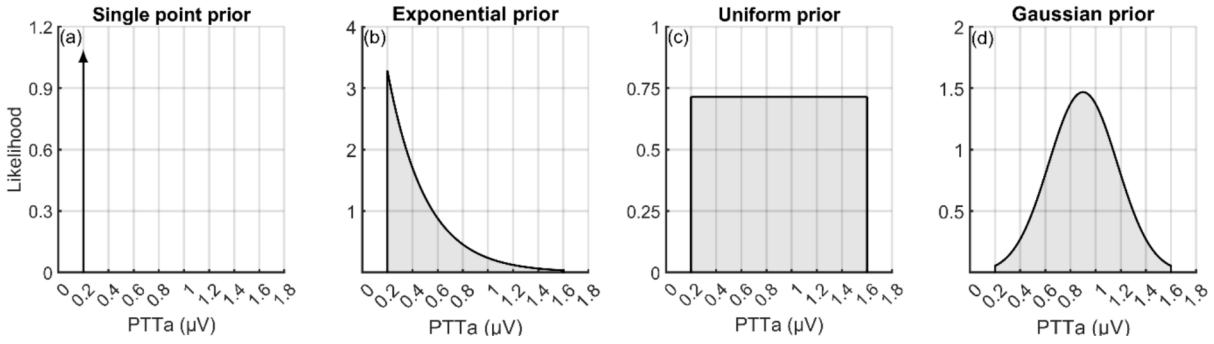

Fig. 2. Four prior distributions for the assumed peak-to-trough amplitude (PTTa) value of the ABR waveform in the data under the alternative hypothesis H_1, including a single point prior, a truncated exponential prior, a uniform prior and a truncated Gaussian prior. These distributions serve as weights when computing the BF value, essentially biasing test sensitivity towards detecting ABR PTTa values that are considered most probable.

$$L(F|H_1) = \int_{A=0.2}^{A=1.6} \varnothing_{NC}(F,\ Q,\ N-Q,\ \lambda_A) \bullet \Pr(A|H_1)\, dA \tag{10}$$

where $\varnothing_{NC}()$ denotes a non-central F distribution, which takes a non-centrality parameter, λ_A, as input:

$$\lambda_A = N \bullet \boldsymbol{\mu}_A \bullet \mathbf{S}^{-1} \bullet \boldsymbol{\mu}_A \tag{11}$$

and where $\boldsymbol{\mu}_A$ is the Q-dimensional vector of voltage means, extracted from the ABR template waveform, rescaled to have a PTTa of A μV. Regarding the integral in Eq. (10), this was computed using the trapezoidal rule with the resolution for A set to 0.01 μV. Note that the integral is repeatedly re-computed each time the analysis is carried out, which was every ~3 seconds in the current work.

3. *Uniform prior*

A uniform prior, shown in Fig. 2, panel (c), was also considered, which was again placed along all ABR PTTa values along the 0.2 to 1.6 μV interval. All ABR PTTa values along this interval were thus deemed equally likely. In this case, $\Pr(A|H_1) = 0.7143$ for all PTTa values, giving an area of one, i.e., 0.7143 x 1.4 = 1. The $L(F|H_1)$ marginal likelihood was computed using Eq. (10) with $\Pr(A|H_1)$ set to the uniform prior.

4. *Truncated Gaussian prior*

Finally, a truncated Gaussian prior, shown in Fig. 2, panel (d), was considered. The most probable PTTa value was now assumed to be the mean value along the 0.2 to 1.6 μV interval, equal to 0.9 μV, and probability densities decreased for PTTa values that were smaller or larger. The standard deviation for this distribution was 0.2718, which led to 99 % of the distribution's area lying along the 0.2 to 1.6 μV interval. Values outside this interval were not considered and their probability density was set to 0. The $L(F|H_1)$ term was again computed using Eq. (10) with $\Pr(A|H_1)$ now set to the truncated Gaussian prior

2.1.4. The prior distribution for the ABR PTTA value under H_0

The null hypothesis of "ABR absent" was always defined using a single A value, namely A = 0 μV. The prior in this case can thus also be considered a single point prior with likelihood values for all A set to zero, except for A = 0.

2.1.5. Sequential testing and critical thresholds

In the current work, the BF value was re-computed every 142 epochs, which corresponds to ~3 seconds intervals when using a 47.17 Hz stimulus rate. For each choice of prior, data were collected and analysed repeatedly until either the $BF > BF_{High}$, or the $BF < BF_{Low}$ criterion was met. Note that no maximum test time was specified. To facilitate a fair comparison with the CGST approach, methods were optimised in simulated data to have a FPR of 0.01 and a TPR of 0.99 (Section 2.3). For the BF approach, this involved optimizing the BF_{High} and BF_{Low} critical thresholds, per choice for prior (Table 1).

2.2. A benchmark to compare against: The CGST approach

To establish a benchmark to compare against, the sequential test strategy from [17] was also included in the assessment, which was previously optimised for ABR detection [11]. This approach uses numerical convolution to find the null distribution of the sequential test statistic, from which the stage-wise critical thresholds for controlling the FPR and true-negative rate (TNR) are derived. As the approach for finding the critical thresholds revolves around convolution, it was coined the Convolutional Group Sequential Test, or CGST.

With the CGST approach, data is analysed in disjoint blocks of observations. Analysing 10,000 epochs with a 5-staged sequential test, for example, might analyse epochs 1–2000 in stage one, epochs 2001–4000 in stage two, etc. At each stage, a test statistic, say T_k, is generated using some statistical test, which is combined with all previously generated test statistics through summation to give the stage k summary statistic, $S_k = \sum_{k=1}^{K} T_k$. In the current work, the T_k test statistics were the log-transformed p values generated by the HT^2 test. The stage-wise summary statistics were thus defined as:

$$S_k = \sum_{k=1}^{K} -2\log(p_k) \tag{12}$$

where p_k is the HT^2-generated p value at stage k. It is worth noting that $-2\log(.)$ is Fisher's transformation [28], which has some desirable properties in terms of test efficiency when combining p values [29]. The S_k test statistics are then evaluated against the stage k critical thresholds: If $S_k < b_k$, then an ABR is deemed absent and H_0 is accepted, whereas if $S_k > a_k$, an ABR is deemed present and H_0 is rejected, else the outcome is deemed ambiguous and the test proceeds to stage k + 1, to a maximum of K stages.

The aim for the CGST approach is to find the a_k and b_k (k = 1, 2, …, K) critical thresholds, such that the stage-wise FPRs and TNRs are controlled. The stage-wise FPRs and TNRs are denoted by α_k and β_k, respectively, and are specified by the user prior to the test. To find a_k and b_k, the CGST first aims to generate the stage-wise null distribution of S_k, say Φ_{S_k}. The only requirement for finding Φ_{S_k} is that the null distributions of the stage-wise T_k test statistics, say Φ_{T_k}, are known. When $T_k = -2\log(p_k)$, then Φ_{T_k} follows a χ^2_2 distribution [28], i.e., a χ^2 distribution with 2 degrees of freedom. For k = 1, $\Phi_{S_1} = \Phi_{T_1} = \chi^2_2$, and for k > 1, the Φ_{S_k} distributions are given by:

$$\Phi_{S_k} = \Phi_{S_{k-1}}^{[b^k, a^k]} * \Phi_{T_k} \tag{13}$$

where * denotes convolution and $\Phi^{[b^k, a^k]}$ indicates that distribution Φ has been truncated to the $[b^k, a^k]$ interval, i.e. all probability densities outside this interval were set to zero. Note that truncation is necessary, as it is not possible to enter stage k with $S_k > a_k$ or $S_k < b_k$, as the sequential test would otherwise already have been stopped. Note also that the area under Φ_{S_k} is reduced by the area of the truncated regions, and thus no longer equals one.

Once the Φ_{S_k} distributions have been generated, finding a_k and b_k is relatively straightforward: a_k is found by solving $\int_{a_k}^{\infty} \Phi_{S_k} = \alpha_k$, and b_k is found by solving $\int_{0}^{b_k} \Phi_{S_k} = \beta_k$. In practice, the ∞ is of course replaced with a sufficiently large value. Additional implementation details and graphical illustrations are also provided in [17].

When using the CGST, various test parameters need to be specified by the user, including (1) the number of stages K, (2) the stage-wise FPRs α_k, (3) the stage-wise TNRs β_k, and (4) the stage-wise ensemble sizes, denoted by N_k. Among these parameters, the choice of K introduces a trade-off between test sensitivity and test time: Larger K implies more frequent data analysis, offering opportunities to stop the test early if a conclusive result is obtained. This helps to keep test time low for the high SNR responses. The trade-off is that the available data is split into smaller chunks. This results in a reduced test sensitivity, potentially prolonging test time for the lower SNR responses.

In previous work [11], the choice for K was evaluated and optimised for ABR detection in both simulations and in subject data: K values

Table 1
The BF_{Low} and BF_{High} critical thresholds, per choice for $\Pr(A|H_1)$ prior, optimised in simulated data to give a False-Positive Rate (FPR) of 0.01 and a true-positive rate (TPR) of 0.99 with reanalysis of accruing data carried out every 3 seconds.

	Single point	Exponential	Uniform	Gaussian
BF_{High}	68.9529	65.15261	39.11059	34.13786
BF_{Low}	0.362637	0.0515147	0.01255055	0.00136935

ranging from 1 to 9 were considered, and in each case, TPRs and FPRs were optimised to be 0.99 and 0.01, respectively, allowing test time to be evaluated as a function of K whilst maintaining equal test specificity and sensitivity. Results show a good test performance across a range of ABR SNRs when using K values of ~5, up to ~9. Based on these results, the current work selected $K = 5$.

For the remaining parameters, the overall β-level was set to 1-α, and the available α and β was split equally across the K stages, giving α_k and β_k values of $\frac{\alpha}{K}$ and $\frac{\beta}{K}$, respectively. To ensure a fair comparison with the BF approach, the overall α-level was optimised in simulations, such that the overall FPR was 0.01, and the ensemble size was optimised to give an overall TPR of 0.99. Data for the optimization was the same simulated data previously used to optimise the BF approach. The optimised ensemble size was 11,625 epochs (or ~246 s), split equally across the K stages, giving N_k values 2325 (for $k = 1, 2, \ldots, 5$). The optimised α-level for obtaining a FPR of 0.01 was 0.0062. The resulting A_k and B_k critical thresholds are shown in Table 2. Note that a_5 and b_5 are equivalent, which means that the test is forced to choose between H_0 and H_1 following stage 5, i.e. H_0 is accepted if $S_5 < b_5$, or rejected if $S_5 > a_5$.

2.3. Evaluating test performance in simulations

The aim for the simulations was thus firstly to first optimise the FPRs and TPRs of the CGST and BF approach (described previously). A second aim was to further evaluate the optimised methods across a range of test conditions. Data comprised coloured noise for emulating the EEG background activity along with a diverse set of ABR templates for simulating a response. These templates differ from the ABR template used to define H_1 in the BF approach. Keeping the assumed ABR template distinct from the true (simulated) ABRs helps to ensure a more fair and realistic evaluation of methods, because the true ABR waveform is usually unknown in practice also.

Simulated EEG background activity: Background activity was simulated using the frequency domain bootstrap (FDB) approach from [30], which takes EEG data as input, and outputs many "surrogate recordings". The surrogates aim to emulate important characteristics of the original data, including signal power, spectral content, and the smoothed envelope of the recording. The main steps of the FDB include (1) rescaling the original recording by its root mean square (RMS) envelope to obtain a recording with a more uniform variance, (2) estimating the power spectral density (PSD) function of the rescaled recording, (3) introducing random variation to obtain many random "surrogate PSD functions", (4) transforming the random surrogate PSD functions back to the time domain using the inverse FFT, and (5) rescaling the surrogates with the previously estimated RMS envelope to restore some non-stationarities in sample variance.

Regarding the RMS envelope, this was estimated by sliding a 200 ms rectangular window over the recording and computing the RMS value at each window location. The resulting RMS envelope was then also smoothed using a 5 Hz low-pass 3rd-order zero-phase (two-pass) Butterworth filter. Regarding the smoothed PSD function, this was estimated using Welch's method [31], which involved (i) segmenting the four second recording into 200 ms windows with 80 % overlap, (ii) applying a Hamming window to each segment to reduce spectral leakage, (iii) taking the FFT of each windowed segment, (iv) transforming the FFT values to PSD values, and (v) averaging the PSD values to obtain a smoothed estimate of the PSD function.

To generate the "random surrogate PSDs", the smoothed PSD function estimates were multiplied (at each frequency) with randomly sampled values from an exponential distribution with a mean of one. The random surrogate PSDs were then transformed to FFT magnitudes, and a random phase value – sampled randomly from a uniform distribution from 0 to 2π – was assigned to each frequency. Finally, the randomised surrogate FFTs were transformed to the time domain using the inverse FFT where they were rescaled by the original smoothed RMS envelope to reintroduce non-stationarities in sample variance.

To further preserve non-stationarities of the original recording, the approach was applied in four second blocks. A new RMS envelope and a new smoothed PSD function was thus computed for each four-second segment of EEG. When simulating longer recordings, surrogates were simulated in four-second blocks and concatenated into a single recording. The original data, from which the surrogates were derived, was approximately 4.5 hours in duration following pre-processing and artefact rejection (Section 2.4.1). For a more detailed description of the FDB approach, including implementation details and illustrative examples, see [30].

ABR templates: As mentioned above, ABRs were simulated using a diverse set of ABR template waveforms. These templates were given by coherently averaged chirp-evoked ABR waveforms, previously recorded from 31 adults, 9 of which had some degree of hearing loss [32]. To qualify as an ABR template, a relatively strict selection criterion was applied to the coherent averages in that the PTTa value should be 10 times larger than the residual background activity. The residual background activity in each coherent average was estimated using the mean absolute difference of its two coherent average replicates: The first replicate was given by the average of the even-numbered epochs, and the second replicate by the odd-numbered epochs. Using such a strict criterion eliminates noisy coherent averages and helps to ensure that a relatively clear (i.e., noise-free) ABR is simulated, rather than background activity. Note that this is a stricter variation of the 3-to-1 rule, which is routinely used by clinicians when detecting ABRs through visual inspection [2]. A total of 30 coherent averages satisfied the selection criterion, which can be seen in the current work's **Supplemental Digital Content**. As described in the sections below, these template waveforms were also rescaled to achieve a certain PTTa value before being added to the no-stimulus surrogates to simulate a response.

2.3.1. Optimizations

As mentioned previously, the critical thresholds for the BF and CGST approach were optimised along with the stage-wise ensemble sizes for the CGST. For the no-stimulus condition, data comprised 500,000 zero-mean surrogates with mean powers ranging from ~1.4 to ~11.9 μV^2, depending on the original EEG recording being emulated. For the stimulus condition, data comprised no-stimulus surrogates with one of the 30 ABR template waveforms (selected at random, per recording) added to each epoch. The PTTa values of the simulated ABRs were varied from 0.2 to 1.6 μV, in steps of 0.01 μV, and a total of 10,000 tests were carried out for each PTTa value, giving a total of 1,410,000 surrogate recordings for the stimulus condition.

2.3.2. Simulations i – Comparisons in sensitivity and test time

Following the initial optimisations, additional simulations were carried out to further evaluate test performance. First, test performance was evaluated for ABR PTTa values ranging from 0.01 to 1.6 μV, in steps of 0.01 μV. These simulations were similar to those used for the optimisation, except that the [0.01, 0.2] μV interval was now also included in the assessment. Note that the [0.01, 0.2] μV interval represents the most challenging cases in the clinic, i.e., cases that tend to be borderline in terms of ABR present/absent. The TPR was now also computed separately, per PTTa value, giving a total of 161 TPRs, each estimated from 10,000 simulated recordings. This contrasts with the preceding optimizations where a single TPR (considering all PTTa values) was

Table 2
The stage-wise critical thresholds when using the Convolutional Group Sequential Test (CGST). These thresholds were optimised, along with the stage-wise ensemble sizes (the N_k values), in simulated data, such that the FPR equalled 0.01 and the TPR equalled 0.99.

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
a_k	0.4432	2.355	5.465	10.12	26.05
b_k	13.39	17.12	20.35	23.31	26.05

estimated.

2.3.3. Simulations II – Evaluating different noise conditions

In the second set of simulations, test performance was evaluated under alternative noise conditions. These simulations were identical to those in Section 2.3.2 above, except that the amplitude of the background activity was now either doubled, or halved. A total of 10,000 tests were again carried out, per PTTa value, and per noise condition. For the no-stimulus condition, the number of tests was increased from 10,000 to 500,000 to obtain a more accurate estimate of the FRP. The 99 % confidence intervals for the expected FPR of $\alpha = 0.01$ were [0.0096, 0.0104], which were found using a binomial distribution constructed from 500,000 "Bernoulli trials" where a "successful Bernoulli trial" was assumed to occur with probability $\alpha = 0.01$.

2.4. Evaluating test performance in subject data

Following the simulations, methods were further evaluated in subject-recorded data, which included no-stimulus EEG recordings [33] and chirp-evoked ABR recordings [32].

2.4.1. No-stimulus EEG data

Recordings of EEG background activity (no stimulus was used) were previously recorded from 17 adults with normal hearing [33] under four test conditions. The current work included recordings from the "sleep" and "still" conditions where subjects were asked to try and sleep (although sleep was not confirmed), or to lie still with their eyes closed but not to fall asleep, respectively. Electrodes (silver–silver chloride, Ag/AgCl) were attached to the left mastoid (active electrode), the right cheek (ground) and the upper forehead (reference), and EEG measurements were made using a Compumedics Neuroscan II EEG amplifier at a sampling rate of 20 kHz. Data were downsampled offline to 5 kHz, after which they were band-pass filtered with a 3rd-order zero-phase (two-pass) Butterworth filter from 30 to 1500 Hz, and restructured into 21.2 ms epochs, corresponding to a (here hypothetical, as no stimuli were applied) stimulus rate of 47.1698 Hz. Artefact rejection was also applied by discarding all epochs with maximum absolute values exceeding 10 μV. After pre-processing and artefact rejection, there were approximately 4.5 hours of data available.

Data analysis: The aim for the no-stimulus data analysis was to estimate the FPR of the methods in subject-recorded data. A challenge, however, is that the BF approach does not specify a maximum test time, and in some cases, there was insufficient data to reach a conclusive test outcome. It was therefore opted to generate a longer set of recordings using the time-domain "moving block bootstrap" approach from [27], which involves resampling (with replacement) blocks of consecutive EEG measurements from the original recordings. The length of the resampled blocks was set to 3 seconds to preserve serial correlation and short-term non-stationarities in the data. For the CGST approach, the total length of the bootstrapped recordings was ~246 seconds (11,625 epochs), whereas for the BF approach this varied per test. A total of 10,000 bootstrapped recordings were constructed and analysed with the BF and CGST methods. It is worth mentioning that the number of independent tests carried out cannot easily be determined due to the resampling with replacement procedure. Confidence intervals for the expected FPR of 0.01 were therefore not constructed.

2.4.2. Chirp-evoked ABR data

Chirp-evoked ABR data were previously recorded from 31 adults (aged 18–70), 9 of whom had some degree of hearing loss [32]. The chirp stimuli included 500, 1000, 2000, and 4000 Hz narrow-band CE-Chirps [34], presented at a rate of 47.1698 Hz at a range of Hearing Levels (dB HL), which were later transformed to dB Sensation Levels (dB SLs) by comparing against the corresponding behavioural hearing thresholds. During the test, subjects were asked to relax with their eyes closed. Electrodes were placed at the vertex (active electrode), the nape of the neck (reference) and mid-forehead (ground), and chirps were presented via a RME Fireface UC soundcard through ER-2 insert phones. Data were recorded using an Interacoustics Eclipse system, and then routed back to Matlab software for recording via the RME sound card with a sampling rate of 48 kHz. Data were subsequently downsampled to 5 kHz, and band-pass filtered from 30-1500 Hz using a 3rd-order zero-phase (two-pass) Butterworth filter. Artefact rejection was also applied by discarding all epochs with maximum values exceeding 20 μV. A more detailed description of the test procedure, including stimulus calibration, is given in [32]. The ensemble sizes for this data ranged from just 1000 epochs, up to 37,000 epochs (median 7000, maximum absolute deviation ~5300). The detection methods were applied to 474 audible (≥0 dB SL) ABR recordings: 124 recordings at ≥ 0–20 dB SL, 169 at ≥ 20–40 dB SL and 181 for ≥ 40 dB SL. After preprocessing and artefact rejection, there was a total of ~22 hours of data available for the analysis. Note that in some cases, there was insufficient data to reach a clear decision, resulting in an ambiguous test outcome. Contrary to the no-stimulus data analysis, recordings were not extended through resampling.

3. Results

This section presents results from the simulations and the subject data analysis.

3.1. Results from simulations i – Comparisons in sensitivity and test time

Results from Simulations I are presented in panels (a) and (b) of Fig. 3: Panel (a) shows the detection rates, per detection method, as a function of the simulated ABR PTTa value, and panel (b) shows the mean test times, per detection method, also as a function of the simulated ABR PTTa value. Note that the PTTa axis in panel (a) was truncated to the [0, 0.3] μV interval to aid visualisation.

Detection rates were similar across methods, with a small advantage being observed for the CGST approach along the [0, 0.2] μV PTTa interval. However, this came at the cost of a relatively large increase in mean test times. When considering the [0, 0.4] μV PTTa interval, for example, the "grand mean test time" (i.e., the mean of the mean test times) was 128.4 seconds for the CGST approach, whereas for the BF approach this was 59.3 (single point prior), 62.3 (exponential prior), 59.3 (uniform prior), and 69.9 seconds (Gaussian prior). For larger PTTa values (i.e., those exceeding 0.4 μV), the mean test time for the CGST approached ~49 s, which was the lowest possible test time as analyses were carried out in blocks of 2325 epochs. This contrasts with the BF approach where the lowest possible test time was ~3 seconds due to the 142 epoch block size. The grand mean test time – considering all PTTa values – was 69.7 seconds for the CGST, whereas for the BF approach this was 28.2 (single point prior), 22.6 (exponential prior), 22.2 (uniform prior), and 24.4 (Gaussian prior) seconds.

Of the four BF priors evaluated, mean test times varied depending on the ABR PTTa value in the data. When considering the [0, 0.3] μV PTTa interval, the lowest mean test times were observed for the single point prior, followed by the exponential prior, the uniform prior, and lastly the Gaussian prior. When evaluating ABR PTTa values larger than ~0.3 μV, the single point prior showed the highest mean test times, whereas the remaining three priors showed similar mean test times and were all lower than the single point prior as well as the CGST. The priors over ABR PTTa values are further considered in the Discussion.

4. Section 3.2. Results from simulations II – Evaluating different noise conditions

Results from Simulations II are presented in panels (c) and (d) of Fig. 3. For the BF approach, results for all four priors showed a similar trend across noise conditions, and to keep this section concise, results are presented for just the single point prior. Results for the remaining

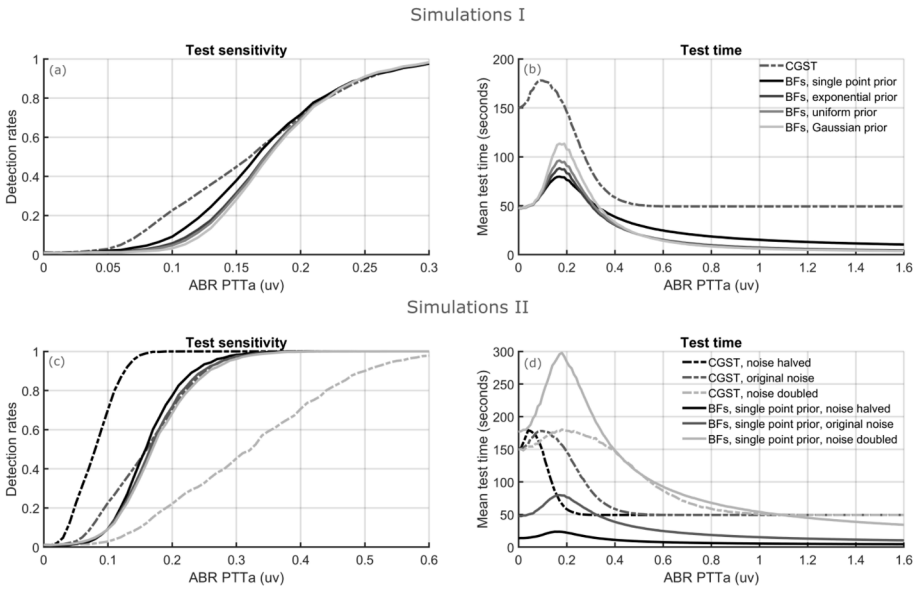

Fig. 3. The detection rates and mean test times of the Bayes Factors (BF) approach and the Convolutional Group Sequential Test (CGST), presented as a function of the simulated ABR PTTa value. Panels (a) and (b) show results from Simulations I and compare detection rates and test times for the CGST and the BF approach with different priors, and panels (c) and (d) show results from Simulations II, which evaluates the effect of changing noise levels in the data. For simulations II, the performance of the BF approach was similar across all four priors. Results were therefore presented for just the single point prior. Note also that the PTTa axis was truncated in panels (a) and (c) for visualization purposes. Further details are provided in the main text.

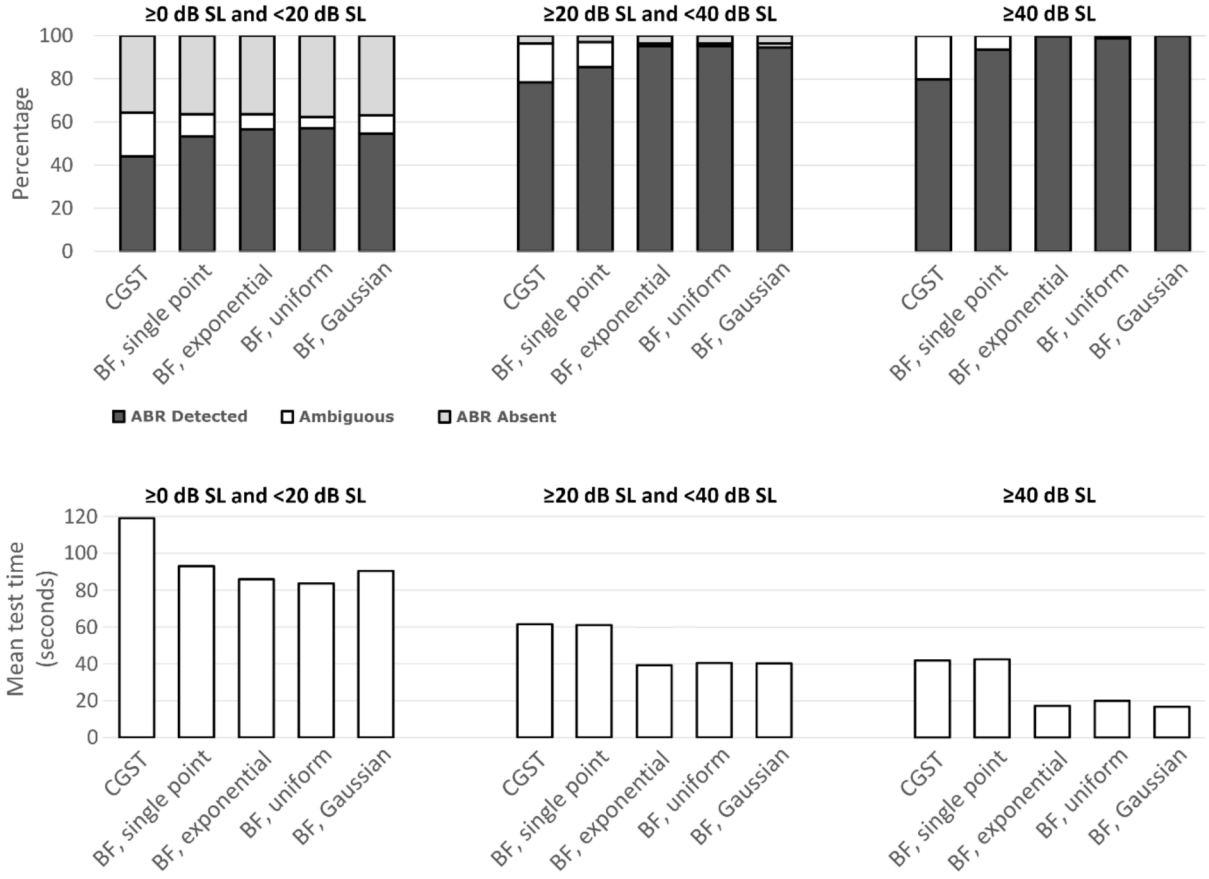

Fig. 4. Results from the subject-recorded chirp-evoked data analysis. The top panels show the percentage of tests where an ABR was deemed present, absent, or "ambiguous", where ambiguous implies that there was insufficient data to reach a clear test outcome. Results are presented as stacked bar plots, per detection method, and per dB sensation level (dB SL) category. The bottom panels show the mean test times, presented as regular bar plots, per detection method, and per dB SL category.

three priors were moved to the **Supplemental Digital Content**.

In terms of detection rates in panel (c), results show that the CGST was impacted considerably by noise levels in the data. For example, when data contained an ABR with a PTTa value of 0.2 μV, detection rates were > 0.999 (noise halved), 0.71 (original noise) and 0.22 (noise doubled). This contrasts with the BF approach where detection rates were relatively stable, i.e., for the 0.2 μV PTTa value, these were 0.77 (noise halved), 0.72 (original noise) and 0.68 (noise doubled). The relatively stable TPRs for the BF approach came at the cost of (or was facilitated by) changes in mean test times: for the 0.2 μV PPTa value, mean test times were 22.2 seconds (noise halved), 76.4 seconds (original noise) and 284.3 seconds (noise doubled).

5. Section 3.3. Results from the chirp-evoked ABR data analysis

Results from the chirp-evoked ABR data analysis are presented in Fig. 4. The top panels show the percentage of tests where an ABR was deemed present, absent, or ambiguous, presented as stacked bar plots, and the bottom panels show the mean test times in seconds. Note that the "ambiguous" category refers to recordings where additional data collection was deemed necessary but was not possible as data were analysed offline. For the higher stimulus levels (SL ≥ 20 dB), the BF approach with an exponential, uniform and/or Gaussian prior showed the highest detection rates as well as the lowest mean test times. For the lower stimulus levels (SL < 20 dB), discrepancies between detection methods were less pronounced, but results suggest slightly higher detection rates and lower mean test times for the BF approach relative to the CGST.

Post-hoc comparisons show that median test times differed significantly across methods for the "≥20 and < 40 dB SL" as well as the "≥ 40 dB SL" categories (global comparisons using Friedman's test, $p < 0.001$), but not for the "≥0 and < 20 dB SL" category (Friedman's test, $p > 0.05$). As a follow-up test, pairwise comparisons were drawn between methods using Fisher's exact test, which confirmed significantly ($p < 0.001$) higher detection rates for the BF approach when using an exponential, a uniform, or a Gaussian prior, relative to both the CGST and the BF approach with a single point prior. The remaining pairwise comparisons revealed no significant differences between methods.

5.1. Test specificity

The estimated FPRs of the detection methods for the simulated data are shown in Table 3. FPRs for the CGST approach fell within the 99 % confidence intervals of the expected FPR of 0.01 for all noise conditions. For the BF approach, FPRs fell within the 99 % confidence for the original noise condition but were significantly lower than α for the noise halved condition, and significantly higher than α for the noise doubled condition. This reduced control over the FPR for the BF approach is not desirable but is the price to be paid for flexibility in test time along with a relatively good control over the TPR (further considered in the Discussion). FPRs for the no-stimulus EEG background activity (also presented in Table 3), however, were approximately controlled as intended, which is reassuring that a good test specificity can be obtained in practice.

With respect to the mean test times for the simulated no-stimulus data (Table 3), it is helpful to first consider the original noise test condition where FPRs were equal across methods, meaning comparisons in mean test times were fair. The mean test times for the BF approach were ~47 seconds under all priors, whereas for the CGST, this was ~149 seconds. The BF approach thus demonstrated a reduced mean test time of almost 70 %. For the remaining noise conditions, the mean test times for the CGST were not impacted, whereas the mean test times for the BF approach varied, ranging from ~13.5 seconds for the noise halved condition to ~177 seconds for the noise doubled condition. For the subject-recorded EEG background activity, mean test times were similar to the mean test times observed for the simulated data under the original noise test condition.

6. Discussion

This study introduced BFs for ABR detection and evaluated its test operating characteristics in simulations and subject-recorded data. Test operating characteristics – specificity, sensitivity and test time – are crucial to the methods performance, but if the approach is to assist examiners in the clinic, then it would ideally also adapt to the needs of the clinician. This involves providing prompt, intuitive and useful feedback for as long as the clinician deems necessary. The BF approach is well-suited in this regard, as it gives frequent feedback (e.g. every ~3 seconds) with no maximum test time specified. As it was designed around the widely used ABR PTTa value, results are interpretable, and potentially more readily integrated into current clinical workflows.

In simulated data, the BF approach also demonstrated reduced mean test times relative to the CGST whilst maintaining equal test specificity and sensitivity: In the no stimulus condition, mean test times were almost 70 % lower for the BF approach relative to the CGST (Table 3). This relatively large reduction in test time is likely because the CGST considers just the null distribution, whereas the BF approach considers both the null and the alternative distribution(s). By considering both distributions, the BF approach can capitalize on recordings with low noise levels, in which case the null and alternative distributions will tend to diverge quickly, leading to more extreme (i.e., smaller or larger) BF values, and ultimately earlier decision-making. For noisy recordings, on the other hand, the null and alternative distributions will diverge more slowly, leading to less extreme BF values, indicating no strong evidence for either hypothesis and hence that additional data collection is needed before a decision can be made.

Thus, when recordings are noisy (relative to the strength of the assumed ABR under H_1), the BF approach tends to reside longer in the indecisive region (the $[BF_{Low}, BF_{High}]$ interval), and automatically prolongs data collection until sufficient evidence has accrued to make a

Table 3
False-positive rates (FPRs) and mean test times (in seconds) from the specificity assessment. Results are presented for the simulated data and for the subject-recorded background activity. For the simulated data, three noise conditions were evaluated, including: (1) the original noise condition, for which methods were optimised to give an FPR of 0.01, (2) the noise halved condition, where the amplitude of the original noise was halved, and (3) the noise doubled condition, where noise amplitude was doubled. Results are presented per noise condition, for both Convolutional Group Sequential Test (CGST) approach and the Bayes Factors (BF) approach with a single point prior (sing.), an exponential prior (exp.), a uniform prior (uni.), or a Gaussian prior (Gauss.).

Methods	**Simulations**						**Subject-recorded background activity**	
	Noise halved		*Original noise*		*Noise doubled*			
	FPR	*Test time*	*FPR*	*Test time*	*FPR*	*Test time*	*FPR*	*Test time*
CGST	0.0104	149.67	0.01	149.34	0.0102	149.2	0.0133	155
BF, sing.	0.0077	13.8	0.0102	47.4	0.0116	176.4	0.0098	41.7
BF, exp.	0.0056	13.5	0.0099	47.1	0.0142	176.6	0.0116	41.8
BF, uni.	0.0041	13.5	0.0099	47.2	0.0177	176.6	0.0099	42
BF, Gauss.	0.0031	13.5	0.0096	47.5	0.0189	177.9	0.0101	42

clear decision regarding which hypothesis (H_0 or H_1) is true. While this helps to control the TPR, it leads to a reduced control over the FPR: for the higher noise levels, time spent in the indecisive region increases, which presents opportunities for random variations to spuriously drive the BF value in the wrong direction, potentially leading to an error and ultimately an increased FPR.

Contrary to the BF approach, the CGST pre-specifies when and how long to analysis data. While this helps to control the FPR, it leads to a reduced control over the TPR. Additionally, the CGST approach only considers the null distribution of the test statistic, i.e., it adopts a frequentist Null Hypothesis Significance Testing (NHST) procedure, which has received increasing levels of critique over the past few decades (e.g., [20,35,36]), particularly in regards to the use of p values: p values represent the probability that data arose under just H_0, and as such, provide a limited assessment of the data as they disregard aspects of statistical power. As a result, p values are effective at providing evidence against H_0, but not in favour of it, meaning they are not well-suited for binary decision tasks such as for ABR detection.

Besides the CGST and BF approach in this work, various additional sequential tests have been proposed in the literature for auditory evoked response detection (e.g., [9,12,15,16,21,37]). Perhaps the most well-known is the Table Testing approach [9,15,16], which uses extensive simulations to find the critical threshold at each test step for controlling the FPR. Like the CGST, table testing has a drawback in that the statistical analysis is specified at the outset, which includes when, how often, and for how long data can be analysed, leading to relatively inflexible test protocols.

6.1. Choice for prior

As the BF is sensitive to the chosen prior [36], the current work evaluated test performance under four different priors. An important factor to consider in this evaluation is that the BF_{Low} and BF_{High} thresholds were optimised, per prior, so that the desired FPR and TPR was obtained. As shown in Table 1, the optimised thresholds depend strongly on the chosen prior, e.g., BF_{Low} was just ~0.0014 for the Gaussian prior, but ~0.363 for the single point prior. This implies that H_0 should be considered 1 / 0.0014 = ~714 times more likely than H_1 before being accepted under a Gaussian prior, whereas under the single point prior, H_0 should be considered 1 / 0.363 = ~2.75 more likely before being accepted.

Broadly speaking, the prior specifies the initial expectations regarding what is plausible, and what is not plausible, and serves as a form of confirmation bias: Evidence that aligns with the initial expectations is deemed important, whereas evidence that contradicts the initial expectations is more readily dismissed. This effect is amplified as the prior distribution becomes less dispersed, or "more informed". The most informed prior in the current work was the single point prior, as this assumed just a single PTTa value of 0.2 μV. Accordingly, mean test times for the single point prior were lowest (relative to the remaining priors) when data conformed with the initial expectation or bias, i.e., when data contained an ABR with a PTTa value equal to (or close to) the assumed 0.2 μV value. The remaining priors were less informed (more dispersed, i.e. higher variance) and considered a wider range of PTTa values at the outset. This led to reduced mean test times (relative to the single point prior) for the larger PTTa values (which were deemed least likely under the single point prior), but higher mean test times (relative to the single point prior) for PTTa values close to 0.2 μV.

Regardless of the choice for prior, mean test time in the current work peaked when data contained an ABR with a PTTa value of just 0.17 μV, which might therefore be considered the most challenging test condition for the BF approach. More generally, cases that are borderline in terms of whether an ABR is present or absent will be challenging, both for the BF approach and for clinicians. It might therefore be preferable to design the detection method for these more challenging scenarios. If so, then results from this work suggest that a single point prior would be the preferred approach. The single point prior has the additional benefit of being relatively easy to implement and interpret.

6.2. Limitations and future work

Perhaps the main limitation for the BF approach in the current work is that its FPR depends on noise levels in the data. When noise levels were changed, the largest FPR variations were observed for the Gaussian prior, and the smallest variations for the single point prior, i.e., for the noise halved condition, FPRs were 0.077 (single point prior) and 0.0031 (Gaussian prior) whereas for the noise doubled condition, these were 0.0116 (single point prior) and 0.0189 (Gaussian prior). This is a concern, as the specificity of the BF approach then depends on test conditions encountered in the clinic, e.g., restless subjects with higher noise levels and longer recordings will tend to result in additional false positives. A potential solution might be to repeatedly re-estimate noise levels from the accruing data and utilize a look-up table to find the correct critical thresholds for controlling the FPR for the estimated noise.

Strictly speaking, the FPR for the BF approach is not just dependent on the noise, but on the SNR. More specifically, it depends on the assumed effect size under the alternative hypothesis relative to the background activity. In the current work, this effect size was defined using an ABR template waveform. Note therefore that if this waveform changes, then the critical thresholds may also need to be adjusted. In future work, a distribution of effect sizes estimated from a set of ABR templates could also be considered. Alternatively, users may prefer to assume the λ parameter directly as this circumvents the need to assume a template.

With respect to the choice for prior, the optimal prior would be both subject- and recording-dependent. Specifically, the optimal prior is a single-point prior, derived from the subject-specific ABR waveform and the recording-specific EEG background activity, both of which, however, are typically unknown. However, even if the optimal prior is usually unattainable, any information about the expected ABR or EEG background activity is valuable and could be used to construct more effective priors, potentially further improving test performance. Future work might therefore aim to identify additional prior information that can be exploited, e.g. whether subjects are expected to have normal hearing or not.

Lastly, the approach was evaluated in subject-recorded chirp-evoked ABR data, but the analysis was not carried out online. Rather it emulated the online data collection procedure with an offline analysis of pre-recorded data. The limitation with this analysis is that there was not always sufficient data to reach an unambiguous test outcome. In future work, the approach should be evaluated online, ideally in a large cohort of infants with suspected hearing loss, as this is expected to be the main target group for this work.

7. Conclusion

Sequentially applied BFs were introduced and evaluated for ABR detection. The approach demonstrated reduced grand mean test times (i.e., the average of all mean test times) of approximately 60–70 % relative to the non-adaptive CGST approach whilst maintaining comparable specificity and sensitivity. The BF approach additionally provides frequent feedback (every ~3 seconds), for as long as the clinician deems necessary, i.e., no maximum test time needs to be specified. A potential limitation, however, is that its FPR depends on noise levels in the data, and further work is needed to obtain a more robust control over test specificity. Care is also needed when selecting the prior, as this greatly impacts on the critical thresholds for response detection. Of the four priors evaluated, the single point prior demonstrated the best performance when detecting ABRs with small amplitudes, which are expected to be the most challenging for clinicians, and thus where clinicians may benefit most from assistance by an objective detection method. Overall,

sequentially applied BFs offer a flexible, intuitive and accurate approach for assisting clinicians with response detection.

CRediT authorship contribution statement

M.A. Chesnaye: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization. **D.M. Simpson:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **J. Schlittenlacher:** Writing – review & editing, Validation, Supervision, Software, Methodology, Funding acquisition. **S. Laugesen:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Funding acquisition. **S.L. Bell:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge Gladys Nijo and Prathyusha Sarika for collecting the chirp-evoked ABR data. 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 for the large-scale simulations. This work was funded by the William Demant Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bspc.2025.107937.

Data availability

Data will be made available on request.

References

[1] Picton, T. W. (2011). Human Auditory Evoked Potentials. San Diego: Plural Publishing Inc.
[2] British Society of Audiology. (2021). Guidelines for the Early Audiological Assessment and Management of Babies Referred from the Newborn Hearing Screening Programme [Online]. Available at: https://www.thebsa.org.uk/resources/ [Accessed: 15/08/2023].
[3] Y.S. Sininger, Audiologic assessment in infants, Curr. Opin. Otolaryngol. Head Neck Surg. 11 (5) (2003) 378–382, https://doi.org/10.1097/00020840-200310000-00012.
[4] R.J. Schmidt, R.T. Sataloff, J. Newman, J.R. Spiegel, D.L. Myers, The sensitivity of auditory brainstem response testing for the diagnosis of acoustic neuromas, Archives of Otolaryngology-Head & Neck Surgery 127 (1) (2001) 19–22, https://doi.org/10.1001/archotol.127.1.19 (PMID: 11177009).
[5] M. Vidler, D. Parkert, Auditory brainstem response threshold estimation: subjective threshold estimation by experienced clinicians in a computer simulation of the clinical test, Int. J. Audiol. 43 (7) (2004) 417–429, https://doi.org/10.1080/14992020400050053 (PMID: 15515641).
[6] M. Zaitoun, S. Cumming, A. Purcell, K. O'Brien, Inter and intra-reader variability in the threshold estimation of auditory brainstem response (ABR) results, Hearing, Balance and Communication 14 (1) (2016) 59–63, https://doi.org/10.3109/21695717.2016.1110957.
[7] C. Elberling, M. Don, Quality estimation of averaged auditory brainstem responses, Scand. Audiol. 13 (3) (1984) 187–197, https://doi.org/10.3109/01050398409043059.
[8] W.H. Martin, J.W. Schwegler, A.L. Gleeson, Y.-B. Shi, New techniques of hearing assessment, Otolaryngol. Clin. North Am. 27 (3) (1994) 487–510, https://doi.org/10.1016/S0030-6665(20)30666-6.
[9] E. Stürzebecher, M. Cebulla, C. Elberling, Automated auditory response detection: statistical problems with repeated testing, Int. J. Audiol. 44 (2) (2005) 110–117, https://doi.org/10.1080/14992020400029228.
[10] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, Objective measures for detecting the auditory brainstem response: comparisons of specificity, sensitivity, and detection time, Int. J. Audiol. 57 (6) (2018) 468–478, https://doi.org/10.1080/14992027.2018.1447697.
[11] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, A group sequential test for ABR detection, Int. J. Audiol. 58 (10) (2019) 618–627, https://doi.org/10.1080/14992027.2019.1625486.
[12] T. Zanotelli, F. Antunes, D.M. Simpson, E. Mazoni Andrade Marçal Mendes, L. B. Felix, Faster automatic ASSR detection using sequential tests, Int. J. Audiol. 59 (8) (2020) 631–639, https://doi.org/10.1080/14992027.2020.1728402.
[13] R.M. McKearney, S.L. Bell, M.A. Chesnaye, D.M. Simpson, Auditory brainstem response detection using machine learning: a comparison with statistical detection methods, Ear Hear. 43 (3) (2022) 949–960, https://doi.org/10.1097/AUD.0000000000001151.
[14] P. Armitage, C.K. McPherson, B.C. Rowe, Repeated significance tests on accumulating data, Journal of the Royal Statistical Society. Series A (general) 132 (2) (1969) 235–244, https://doi.org/10.2307/2343787.
[15] E. Stürzebecher, M. Cebulla, Automated auditory response detection: improvement of the statistical test strategy, Int. J. Audiol. 52 (12) (2013) 861–864, https://doi.org/10.3109/14992027.2013.822995.
[16] M. Cebulla, E. Stürzebecher, Automated auditory response detection: further improvement of the statistical test strategy by using progressive test steps of iteration, Int. J. Audiol. 54 (8) (2015) 568–572, https://doi.org/10.3109/14992027.2015.1017659.
[17] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, The convolutional group sequential test: reducing test time for evoked potentials, IEEE Trans. Biomed. Eng. 67 (3) (2020) 697–705, https://doi.org/10.1109/TBME.2019.2919696.
[18] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, Constructing group sequential tests with data monitoring, Biomed. Signal Process. Control 94 (2024) 106278, https://doi.org/10.1016/j.bspc.2024.106278.
[19] R.E. Kass, A.E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (430) (1995) 773–795, https://doi.org/10.1080/01621459.1995.10476572.
[20] F.D. Schönbrodt, E.J. Wagenmakers, M. Zehetleitner, M. Perugini, Sequential hypothesis testing with Bayes factors: efficiently testing mean differences, Psychol. Methods 22 (2) (2017) 322–339, https://doi.org/10.1037/met0000061.
[21] L. Wang, E. Noordanus, A.J. Van Opstal, Towards real-time detection of auditory steady-state responses: a comparative study, IEEE Access 9 (2021) 108975–108991, https://doi.org/10.1109/ACCESS.2021.3100157.
[22] H. Hotelling, The generalization of student's ratio, Ann. Math. Stat. 2 (1931) 360–378, https://doi.org/10.1214/aoms/1177732979.
[23] Rencher, A. C. 2001. Methods of Multivariate Analysis. 2nd ed. 118. Hoboken, NJ: John Wiley & Sons, Inc.
[24] M. Bilodeau, D. Brenner, Theory of Multivariate Statistics, Springer, New York, 1999, p. 100.
[25] Li, J., Qiu, Y., Li, L. (2017). A Neighborhood-Assisted Hotelling's T2 Test for High-Dimensional Means. Eprint arXiv:1712.01798, (2017).
[26] J.K. Nousak, D.R. Stapells, Auditory brainstem and middle latency responses to 1 kHz tones in noise-masked normally-hearing and sensorineurally hearing-impaired adults, Int. J. Audiol. 44 (6) (2005) 331–344, https://doi.org/10.1080/14992020500060891.
[27] J. Lv, D.M. Simpson, S.L. Bell, Objective detection of evoked potentials using a bootstrap technique, Med. Eng. Phys. 29 (2) (2007) 191–198, https://doi.org/10.1016/j.medengphy.2006.03.001.
[28] R.A. Fisher, Statistical methods for research workers, 11th ed., Oliver and Boyd, Edinburgh, U.K., 1932.
[29] R.C. Littell, J.L. Folks, Asymptotic optimality of Fisher's method of combining independent tests, J. Am. Stat. Assoc. 66 (336) (1971) 802–806, https://doi.org/10.2307/2284230.
[30] M.A. Chesnaye, S.L. Bell, J.M. Harte, D.M. Simpson, Controlling test specificity for auditory evoked response detection using a frequency domain bootstrap, J. Neurosci. Methods 363 (2021) 109352, https://doi.org/10.1016/j.jneumeth.2021.109352.
[31] P. Welch, The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modifiied periodograms, IEEE Trans. Audio Electroacoust. 15 (1967) 70–73, https://doi.org/10.1109/TAU.1967.1161901.
[32] M.A. Chesnaye, D.M. Simpson, J. Schlittenlacher, S. Laugesen, S.L. Bell, Audiogram estimation performance using auditory evoked potentials and Gaussian processes, Ear Hear. 46 (1) (2025) 230–241, https://doi.org/10.1097/AUD.0000000000001570.
[33] S.M.K. Madsen, J.M. Harte, C. Elberling, T. Dau, Accuracy of averaged auditory brainstem response amplitude and latency estimates, Int. J. Audiol. 57 (5) (2018) 345–353, https://doi.org/10.1080/14992027.2017.1381770.
[34] C. Elberling, M. Don, A direct approach for the design of chirp stimuli used for the recording of auditory brainstem responses, J. Acoust. Soc. Am. 128 (5) (2010) 2955, https://doi.org/10.1121/1.3489111.

[35] F.D. Schönbrodt, E.J. Wagenmakers, Bayes factor design analysis: planning for compelling evidence, Psychon. Bull. Rev. 25 (1) (2018) 128–142, https://doi.org/10.3758/s13423-017-1230-y (PMID: 28251595).
[36] D.W. Heck, U. Boehm, F. Böing-Messing, P.C. Bürkner, K. Derks, Z. Dienes, Q. Fu, X. Gu, D. Karimova, H.A.L. Kiers, I. Klugkist, R.M. Kuiper, M.D. Lee, R. Leenders, H. J. Leplaa, M. Linde, A. Ly, M. Meijerink-Bosman, M. Moerbeek, J. Mulder, B. Palfi, F.D. Schönbrodt, J.N. Tendeiro, D. van den Bergh, C.J. Van Lissa, D. van Ravenzwaaij, W. Vanpaemel, E.J. Wagenmakers, D.R. Williams, M. Zondervan-Zwijnenburg, H. Hoijtink, A review of applications of the Bayes factor in psychological research, Psychol. Methods 28 (3) (2023) 558–579, https://doi.org/10.1037/met0000454.
[37] F. Bardy, B. Van Dun, M. Seeto, H. Dillon, Automated cortical auditory response detection strategy, Int. J. Audiol. 59 (11) (2020) 835–842, https://doi.org/10.1080/14992027.2020.1767808.