Taylor & Francis
Taylor & Francis Group

Original Article

International Journal of Audiology 2005; 44:110–117

*Ekkehard Stürzebecher**
*Mario Cebulla**
Claus Elberling[†]

*ENT Clinic, Faculty of Medicine,
Johann Wolfgang Goethe-University
Frankfurt am Main, Germany
[†]Oticon A/S Research Centre,
Eriksholm, Snekkersten, Denmark

Automated auditory response detection: Statistical problems with repeated testing

Evaluación repetida en la detección de respuestas auditivas

Key Words

Automated response detection
Repeated testing
Dependent samples
Bonferroni correction

Abstract

Sequential application of a statistical test is usually applied in an automated auditory response detection algorithm. The sequential test strategy is very time-efficient but increases the probability of a false rejection of the null-hypothesis. For this reason, it is necessary to correct the critical test value. However, the well-known Bonferroni correction leads to an over-correction when dealing with dependent or partly dependent data. The objective of the study reported here was to develop a method to determine the critical test value for the sequential testing of dependent data. Extensive Monte Carlo simulations were used to develop this method. The simulation results were reviewed and the benefit of the suggested method, in comparison to the Bonferroni correction, was shown using a large sample of real amplitude modulation following response data. The detection rate determined for these data and the ROC curve demonstrate the advantage of using the method suggested here.

Sumario

La aplicación secuencial de pruebas estadísticas se usa normalmente con algoritmos automatizados para la detección de respuestas auditivas. La estrategia de pruebas secuenciales es muy eficiente en tiempo pero incrementa la probabilidad de falso rechazo de la hipótesis nula. Por esta razón, es necesario corregir el (los) valor(es) crítico(s) de la prueba. No obstante, la bien conocida corrección de Bonferroni lleva a una sobre-corrección cuando se trata de datos dependientes o parcialmente dependientes. El objetivo de este estudio fue desarrollar un método para determinar el valor crítico de la prueba, para la evaluación secuencial de datos dependientes. Se usaron amplias simulaciones Monte Carlo para desarrollar este método. Se revisaron los resultados de la simulación y se mostró el beneficio del método sugerido, en comparación con la corrección Bonferroni, con el uso de amplias muestras de modulación real de amplitud, siguiendo los datos de respuesta. La tasa de detección determinada por estos datos y la curva ROC demuestran las ventajas del uso del método que aquí se sugiere.

Introduction

When assessing the hearing threshold by means of auditory steady-state responses (ASSR), or screening the hearing in newborns by means of otoacoustic emissions (OAE) or auditory brainstem responses (ABR), a stepwise (sequential) application of the statistical decision criterion which is carried out simultaneously with the data acquisition, is usually preferred. This means that the statistical test is applied for the first time, as soon as a predefined minimum number of stimulus-related epochs are available. If no response is detected by this first test, the sample will be extended by a given number of epochs and the test will be carried out again. This procedure is repeated until a response is detected, a predefined maximum number of epochs have been reached or a predefined maximum examination time has expired (see e.g. the general discussion by Don & Elberling, 1996). This successive procedure is more efficient than using a fixed sample size and a single statistical testing. If, for example, a small sample size is chosen, small responses may not be detected and if a large sample size is chosen, time might be wasted detecting responses of a high amplitude that already could have been detected using a considerably smaller sample size. However, a clear disadvantage of the sequential test strategy is that multiple statistical testing increases the probability of false rejection of the null-hypothesis (Lütkenhöner, 1991). Therefore, in the case of repeated testing, a correction of the significance level is necessary (Hochberg & Tamhane, 1987). If a significance level of p is to be ensured in n tests using independent data, testing has to be performed at a significance level of $p' = p/n$ in accordance with Bonferroni's rule. This has the advantage that the required probability p for false rejection of the null-hypothesis (false pass in hearing screening) is fulfilled. However, the probability of false rejection of the alternative hypothesis, H_A, becomes inevitably higher as a consequence of the increase in the critical test value because an existing response may no longer be detected. Up to now, we have used the Bonferroni correction (Cebulla et al, 2000), and the necessity of using this correction when using a successive test strategy for hearing screening in newborns has also been emphasized by Picton et al (2002). However, similar comments have not been made in other publications dealing with objective determination of hearing threshold and hearing screening in newborns. This leaves the impression that this specific problem may have been overlooked during the practical application of the different statistical test procedures with commercial devices.

The above-mentioned Bonferroni correction only applies if multiple tests use independent data. With dependent data the

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.1080/14992020400029228
© 2005 British Society of Audiology, International Society of Audiology, and Nordic Audiological Society

Received:
July 11, 2003
Accepted:
May 3, 2004

Ekkehard Stürzebecher
Klinikum der J. W. Goethe-Universität, Klinik für Hals-Nasen-Ohrenheilkunde
Theodor-Stern-Kai 7, D-60590 Frankfurt am Main, Germany
E-mail: stuerzebecher@em.uni-frankfurt.de

Bonferroni correction is too conservative (Hochberg & Tamhane, 1987), i.e. a smaller level of p' than necessary is chosen to fulfil the given significance level p. These facts apply to the sequential procedures used for the objective determination of hearing threshold and hearing screening in newborns. The individual tests are not based on independent data because the current sample consists of the previous one with the addition of one or more new epochs. A correction in accordance with Bonferroni is therefore not optimal for sequential testing of such samples. By using the Bonferroni correction, the probability for false rejection of the null-hypothesis is reduced beyond the extent required and, furthermore, the time required for the response detection is prolonged.

The objective of the study reported here was to demonstrate over-correction when applying the Bonferroni correction and to develop a method, to determine the critical test values for repeated testing of dependent samples, independent of the applied statistical test procedure. The development of the method was carried out by means of extensive Monte Carlo simulations. The simulation results were reviewed and the benefit of the suggested method compared to the Bonferroni correction was shown on the basis of a large sample of real Amplitude-Modulation Following Response (AMFR) data.

Methods

The method used for the Monte Carlo simulations has already been described in detail (Stürzebecher et al, 1999): Pairs of Gaussian distributed random numbers were generated using the Box-Müller method (Press et al, 1986). These pairs could be regarded as the components a_i and b_i of a vector $\boldsymbol{V}_i$ in Cartesian coordinates. The phase angle is calculated in accordance with $\varphi_i = \arctan(b_i/a_i)$ and the vector length is $|\boldsymbol{V}_i| = (a_i^2 + b_i^2)^{1/2}$.

The analyses were carried out using the modified Rayleigh test as a model (Moore, 1980). The modified Rayleigh test is a test in the frequency domain that evaluates the phase and amplitude of a harmonic (usually the first harmonic) of the spectrum of an 'auditory steady-state response' (ASSR). In a previous article (Cebulla et al, 2001), we have shown that the modified Rayleigh test is one of the most powerful tests, therefore this test was chosen for the present study. However, any other known statistical test for response detection could have been used.

The test statistics of the modified Rayleigh test are

$$R_m^* = \frac{R_m}{n\sqrt{n}}$$

with

$$R_m = \sqrt{C_m^2 + S_m^2};$$

$$C_m = \frac{1}{n}\sum_{i=1}^{n} r_i \cos\varphi_i; \quad S_m = \frac{1}{n}\sum_{i=1}^{n} r_i \sin\varphi_i$$

φ_i – spectral phase angle;
r_i – rank number of $\boldsymbol{V}_i$; $1 \leq r_i \leq n$;
$\boldsymbol{V}_i$ – spectral amplitude;
n – sample size.

In an earlier study concerning the objective detection of the AMFR carried out by means of different one-sample tests (Cebulla et al, 2001), a sample size with the maximum number of 100 epochs was found to be well suited. This forms the basis for the simulations carried out within the scope of the present study. 8×10^6 data sets each consisting of 100 pairs of random numbers (sample size 100) were generated. The high number of 8×10^6 data sets was chosen to make sure that reliable estimates of the two null hypotheses were obtained.

Each data set was tested using the modified Rayleigh test, starting with a sample size of 10. This sample was stepwise extended by one pair of random numbers (increment 1) until all 100 pairs were included. In this manner, 91 test results (91 test values) were achieved for each data set. Two different frequency distributions H_01 and H_02, each with 1000 classes of test values, were constructed from the test values of the 8×10^6 data sets in the following way.

Frequency distribution H_01 and H_02

Each complete data set was described by two parameters: its final test value and its maximum test value. The frequency distribution H_01 was formed by the 8×10^6 final test values, whereas the frequency distribution H_02 was formed by the 8×10^6 maximum test values.

The critical test values for $p = 0.01$ ($\alpha = 1\%$) were calculated on the basis of the two distributions and the Bonferroni correction of the critical test value for $n = 91$, determined by means of H_01, was carried out.

The same calculations as described above for $n = 91$, were also carried out for $n = 16$ because many authors use a markedly lower number of test steps in automatic ASSR detection, frequently $n = 16$ (Champlin, 1992; Dobie & Wilson, 1993; Lins & Picton, 1995; Valdes et al, 1997).

H_01 is independent of the number of test steps because only those test values resulting from the single testing of the entire sample are included in the distribution. To distinguish between the two different H_02 distributions, the respective distributions and critical test values are assigned the indices 91 or 16 corresponding to the number of test steps (i.e. H_02_{91} or H_02_{16}).

The evaluation of the simulation results was carried out on the basis of a large sample of real Amplitude-Modulation Following Response (AMFR) data (Cebulla et al, 2001). The sample contained the raw data of 1484 stored AMFR recordings, each with a length of 102.4 s. The data were recorded from 57 male and female adults, aged between 20 and 64 years. 46 subjects had normal hearing with threshold of 10 dB HL or better at 500 – 4000 Hz. 11 subjects had a sensorineural hearing loss of at least 30 dB, but no more than 65 dB for at least one of the four frequencies 500, 1000, 2000, and 4000 Hz. The subjects reclined comfortably on an examination couch in a soundproof and electrically shielded room. They were asked to relax and if possible to sleep during the examination. For the normally hearing subjects, the stimulus level was 30 dB nHL. For the patients with hearing loss, the stimulus level was 30 dB SL. The low stimulus level was chosen because differences in test performance are more apparent for stimulus levels near threshold. (For further details see Cebulla et al, 2001.)

The AMFR is composed by response energy located at several harmonics in the frequency domain, where the frequency of the first harmonic corresponds to the modulation frequency (in this case approximately 90 Hz). The frequencies of the higher harmonics are integer multiples of the modulation frequency. Each set of data was divided into 100 epochs with a length of approximately 1 s. The epochs were transformed to the frequency domain by means of a Discrete Fourier Transformation (DTF). The frequency resolution amounts to about 1 Hz as a result of the chosen epoch length.

These data were first used to investigate whether the suggested solution leads to a statistically correct result and next, to estimate the gain in test performance.

Evaluation of the statistical correctness of the suggested procedure

As a consequence of the selected epoch length of approximately 1 s and the modulation frequency of approximately 90 Hz, 89 spectral components that only contained background noise were positioned between the spectral components of the harmonics of the response. 25 000 of these data sets, each consisting of 100 spectral noise components, were subjected to the modified Rayleigh test. The distributions $H_0 1$ and $H_0 2_{91}$ were constructed with these 25 000 test values in the same manner as described above for the simulated data. Additionally, each data set was tested once using the critical test value for $\alpha(H_0 1) = 1\%$, but in 91 or 16 test steps with the critical test values for $\alpha(H_0 2_{91}) = 1\%$ and $\alpha(H_0 2_{16}) = 1\%$ and with the test values of $n = 91$ and $n = 16$, corrected in accordance with Bonferroni.

Evaluation of the gain in test performance

The modified Rayleigh test was applied to the spectral component of the first harmonic (fundamental frequency) of the AMFR, at approximately 90 Hz, to determine the gain in test performance. The analysis was based on the following critical test values:

- critical test values for $\alpha(H_0 1) = 1\%$ for $n = 91$ and $n = 16$, corrected in accordance with Bonferroni;
- critical test values for $\alpha(H_0 2_{91}) = 1\%$ and $\alpha(H_0 2_{16}) = 1\%$.

Testing of each of the 1484 AMFR data sets was carried out using both the critical test values corrected in accordance with Bonferroni and those critical test values determined from $H_0 2_{91}$ and $H_0 2_{16}$, using 91 or 16 test steps.

Performance was compared on the basis of the detection rate determined for the AMFR data and on the basis of the mean detection time. The mean detection time is the mean value of the time required for the detection of individual responses; it was calculated from the number of epochs required for response detection.

In addition, receiver-operating characteristics, ROC-curves (Green & Swets, 1966; Howell, 1987; Lütkenhöner, 1991) were constructed from the probability density functions of H_0 estimated by the described Monte Carlo simulations and from the distribution of the test values calculated for the 1484 AMFR-data sets (maximum of 91 test steps) which acts as an estimate of the alternative hypothesis, H_A. A ROC-curve relates the probability of true acceptance of the alternative hypothesis, H_A,(HITS) to the probability of false rejection of the null-hypothesis, H_0, (FALSE ALARMS), and allows a comprehensive assessment of the detection performance of a statistical test procedure. For the construction of the ROC-curve for the Bonferroni correction, $H_0 1$ was re-plotted so all original α-values now corresponded to values that were 91 times smaller; for example, the $H_0 1$-point for $\alpha = 1\%$ was moved to the position for $\alpha = 0.011\%$ (the Bonferroni-corrected test level).

Results and discussion

Simulated data

Figure 1 shows the frequency distributions $H_0 1$ and $H_0 2_{91}$ derived from the 8×10^6 simulations with the modified Rayleigh test. (For presentation purposes, 10 of each of the 1000 classes were consolidated into a total of 100 classes.) The corresponding distributions constructed from the 25 000 AMFR data sets containing only background noise are shown for comparison in Figure 1. 25 000 data sets were obviously too few to produce a

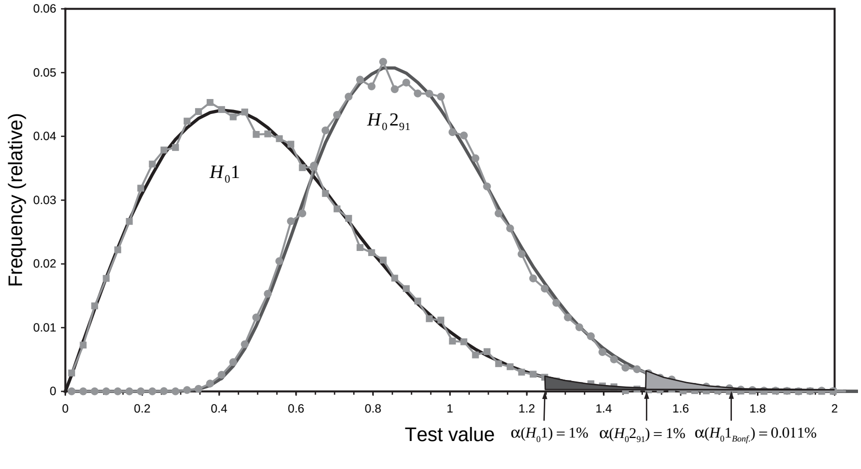

Figure 1. Simulated frequency distributions $H_0 1$ and $H_0 2_{91}$ with the critical test values for $\alpha(H_0 1) = 1\%$, $\alpha(H_0 2) = 1\%$, and $\alpha(H_0 1_{\mathrm{Bonf.}}) = 0.011\%$ indicated by arrows. The 8×10^6 final test values that resulted from the test of each data set using all 100 pairs of random numbers were used for the frequency distribution $H_0 1$. To form the frequency distribution $H_0 2_{91}$, only the 8×10^6 maximum test values that resulted from the 91 tests of each data set were used. The corresponding distributions constructed from the 25 000 AMFR data sets containing only background noise are shown for comparison in Figure 1.

sufficiently smooth distribution. However, the simulated distributions fit the H_0-distributions of the real AMFR data very well. This indicates that the simulated data lead to reliable estimates of the probability density functions of the null-hypotheses for the different test situations. With simulations, a great number of samples (here 8×10^6 samples with 100 pairs of random numbers each) is more easily attainable than with real data. Such a large number of samples is necessary to get a high resolution especially in the extreme tail of the distributions. This is important for specifying the critical test values at a very low α-level (for instance $\alpha = 0.01\%$) which is necessary for screening applications.

Figure 2 shows all three simulated frequency distributions: $H_0 1$, $H_0 2_{91}$ and $H_0 2_{16}$. $H_0 1$ is an estimate of the null-hypothesis when using a single testing of a sample. Because of the initial considerations, it is assumed that $H_0 2_{91}$ is an estimation of the null-hypothesis in a case where 91 successive tests are carried out using dependent samples. The highest of the 91 test values of each of the simulated 8×10^6 samples was used in each case to construct the distribution $H_0 2_{91}$. This maximum test value is the test value that can lead to a false rejection of the null-hypothesis in the case of sequential testing. The distribution of these maximum test values indicates the worst cases in successive testing of dependent samples. The critical test value for a predefined error probability in multiple tests of dependent samples can also be directly derived from the distribution $H_0 2_n$, in the same way as in the case with single tests from $H_0 1$ (without additional correction). This is shown for 91 sequential test steps in Figure 1. An error probability of $\alpha = 1\%$ was chosen as the basis for comparing the results. The critical test value from the distribution $H_0 1$ was chosen for the single test of a sample with the modified Rayleigh test. The critical test value for $\alpha(H_0 1) = 1\%$ is 1.25. The Bonferroni correction for 91 test steps leads to an error probability of $\alpha(H_0 1_{\text{Bonf.}}) = \alpha(H_0 1)/91 = 0.011\%$. The corresponding critical test value from $H_0 1$ amounts to 1.73. In contrast, the critical test value resulting from $H_0 2_{91}$, for the 91-fold successive testing of dependent samples with $\alpha(H_0 2) = 1\%$, amounts to only 1.51.

The critical test value for the nonrecurring testing is the smallest among the 3 critical test values specified. Thus, with a Type I error (false rejection of the null-hypothesis H_0 when $\alpha = 1\%$) that is identical for all three variants, the Type II error (false rejection of the unknown alternative hypothesis $H_{A,}$ located to the right of H_0 in Figure 1) is the smallest in testing once. Consequently, the sensitivity-index (a measure of distance between the distribution of the null-hypothesis and the alternative hypothesis) is highest when a sample is subjected to single testing.

The disadvantage represented by a slight reduction in the sensitivity index, brought about by the required increase of the critical test value, has to be accepted if one wants to benefit from the advantages of the sequential testing described above. However, the loss in sensitivity index is significant when dependent samples are tested using a critical test value corrected in accordance with Bonferroni's rule.

The aim of this study was to restrict the loss in sensitivity-index to the necessary extent when testing dependent samples. The critical test value resulting from $H_0 2_{91}$ for the successive 91-fold testing of dependent samples is markedly lower than the critical test value from $H_0 1$ corrected in accordance with Bonferroni. This leads to a higher sensitivity index than that achieved with the critical test values corrected in accordance with Bonferroni.

AMFR data – test application on the background noise spectral components

Initially, we had to examine whether the suggested method provided statistically correct test values. For this purpose, the modified Rayleigh test was applied to the spectral components from 25 000 AMFR data sets that only contained biological background noise. Each set of data comprising 100 background-noise spectral components was tested once using the critical test value for $\alpha(H_0 1) = 1\%$. Furthermore, the data sets were tested in 91 or 16 test steps using the critical test values for $\alpha(H_0 2_{91}) = 1\%$ and $\alpha(H_0 2_{16}) = 1\%$ and the test values for $n = 91$ and $n = 16$ corrected in accordance with Bonferroni.

In all test conditions, accidental response detection can be expected in a particular number of data sets, despite the fact that the background-noise spectral components contain no stimulus-related power. In single testing of the sets of data with the critical

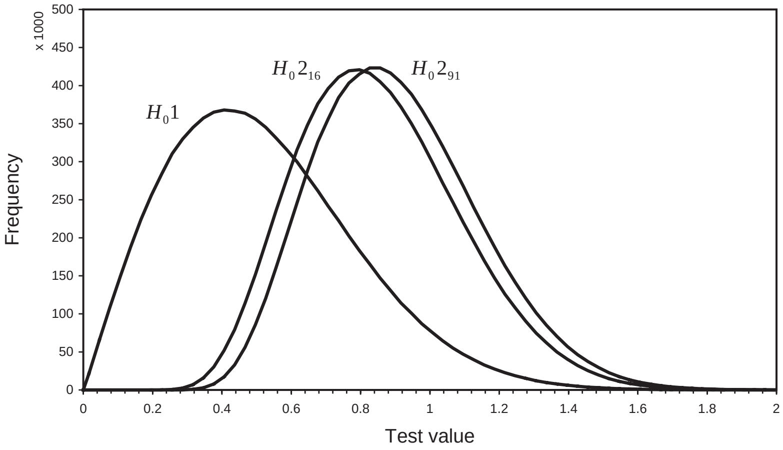

Figure 2. Simulated frequency distributions $H_0 1$, $H_0 2_{16}$, and $H_0 2_{91}$. To form the frequency distribution $H_0 2_{16}$, only the 8×10^6 maximum test values that resulted from the 16 tests of each data set were used.

test value for $\alpha(H_01)=1\%$, a false rejection of H_0 must be expected to appear in approximately 1% of the data sets. If the critical test values for the multiple testing of dependent samples are correct, an accidental rejection of the null-hypothesis will also take place in approximately 1% of the data sets when testing the background noise spectral component 91 or 16 times using the critical test value for $\alpha(H_02)=1\%$. In contrast, the percentage of an accidental rejection of the null-hypothesis should be markedly below 1% if the correction of the critical test values in accordance with Bonferroni is actually too conservative.

A response was erroneously detected in 237 data sets (false rejection of the null-hypothesis) during the single testing of the background noise samples. This corresponds to an error rate of 0.95% (237/25,000) – very close to the predefined error probability of $\alpha=1\%$. Table 1 contains the values that resulted from the multiple testing of the samples. The critical test values corrected in accordance with Bonferroni, results in an error rate of 0.29% using 16 test steps and an error rate of 0.13% using 91 test steps. These values are considerably below the predefined error probability of 1%. Thus, it is confirmed that the procedure carried out in accordance with Bonferroni actually leads to over-correction in the case of dependent samples. The error rate for 16 test steps is closer to the predefined error probability of 1% than the value for 91 test steps. With a decreasing number of test steps, any over-correction caused by applying the Bonferroni correction obviously becomes smaller.

Using the critical test values for $\alpha=1\%$ determined from H_02_{91} or H_02_{16}, an error rate of 0.95% results for 91 test steps and 1.02% for 16 test steps. Both values correspond approximately to 1%, i.e. the predefined error probability. The difference between the error rates for the H_02 condition and for the Bonferroni correction is highly significant ($p<0.0001$, Fishers exact test) for 91 test steps and 16 test steps as well. These results demonstrate that the procedure suggested for determining the critical test values for the multiple testing of dependent samples provides correct critical test values.

In Figure 3, the critical test values ($\alpha=1\%$) are plotted as a function of the number of test steps for the Bonferroni correction as well as for the new procedure suggested here. The critical test value corrected according to Bonferroni's rule increases steadily with increasing number of test steps. In contrast, the critical test value assessed from H_02 shows a significant increase only up to about 30 test steps and a further increment in the number of test steps only leads to a marginal increase of the critical test value. The reason for this is that the amount of dependency increases with each new epoch that is added to the sample, and at the same time the amount of change in the statistical characteristics of the sample decreases with increasing sample size.

Table 1. Error rates (probability of a false rejection of the null-hypothesis) achieved when testing the background-noise spectral components of the AMFR data set using the critical test values determined from H_02_{16} and H_02_{91}, as well as the critical test values corrected in accordance with Bonferroni for $n=16$ or $n=91$ from H_01. The selected error probability is $\alpha=1\%$. The single testing of each data set with the critical test value determined for $\alpha=1\%$ from H_01 results in an observed error rate of 0.95%.

	H_02	*Bonferroni*
16 test steps	1.02%	0.29%
91 test steps	0.95%	0.13%

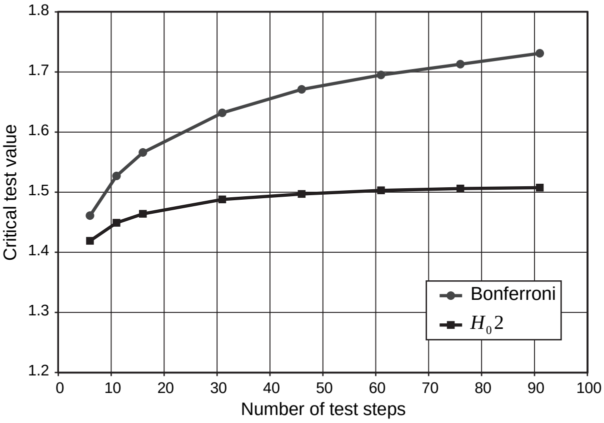

Figure 3. Critical test values ($\alpha=1\%$) plotted as a function of the number of test steps for the Bonferroni correction as well as for the new procedure using the estimated distribution H_02.

AMFR data – test application on the spectral response component

When the 1484 data sets consisting of 100 epochs are tested once using the critical test value for $\alpha=1\%$ determined from H_01, a detection rate of 87.1% is achieved. This value is slightly lower than the detection rate of 88.6% stated in a previous publication for the same set of data with the application of the modified Rayleigh test (Cebulla et al, 2001). The reason for this is that the critical test value (1.240) determined for H_01 using a smaller number of Monte Carlo simulations was slightly lower than the presently-determined critical test value (1.248), where H_01 was estimated with 8×10^6 simulated test values. Using the critical test values determined from H_02_{16} and H_02_{91} as well as the critical test values from H_01 for $n=16$ or $n=91$, corrected in accordance with Bonferroni, the detection rate is given in Table 2 and the mean detection time in Table 3, for testing the spectral response component (first harmonic) of the AMFR data. The error probability was $\alpha=1\%$. In principle, it is not to be expected that the detection rate determined in single testing is obtained in multiple testing because the critical test values for multiple testing are higher than the critical test value for single testing. A detection rate of 80.8% was achieved for 16 test steps (H_02_{16}) and a detection rate of 79.8% was achieved for 91 test steps (H_02_{91}). In accordance with expectations, the detection rate for 16 test steps is slightly higher than that for 91 test steps because the critical test value for 16 test steps is slightly lower. In return,

Table 2. Detection rate for the testing of the spectral response components (first harmonic) of the AMFR data set, using the critical test values determined from H_02_{16} and H_02_{91} and the critical test values corrected in accordance with Bonferroni, determined for $n=16$ or $n=91$ from H_01. The chosen error probability is $\alpha=1\%$.

	H_02	*Bonferroni*
16 test steps	80.8%	74.9%
91 test steps	79.8%	67.6%

Table 3. Mean detection time ± standard deviation for the testing of the spectral response component (first harmonic) of the AMFR data set using the critical test values determined from H_02_{16} and H_02_{91} and the critical test values corrected in accordance with Bonferroni, determined for $n = 16$ or $n = 91$ from H_01. The selected error probability is $\alpha = 1\%$.

	H_02	*Bonferroni*
16 test steps	42.1 ± 26.4 s	44.8 ± 25.8 s
91 test steps	39.8 ± 26.0 s	45.7 ± 24.8 s

the time effectiveness is less favourable for 16 test steps. This is because, in the case of 100 epochs with a length of approximately one second, 16 test steps produce a step-width of 5 epochs if the first test is carried out with 10 epochs. Here, the detection time is determined with a time frame of approximately 5 s. In the case of 91 test steps, the frame corresponds to the epoch length of 1 s.

The detection rates (16 test steps: 74.9%; 91 test steps: 67.6%) determined with the critical test values corrected in accordance with Bonferroni are lower than the detection rates for the critical test values determined from H_02_{91} and H_02_{16}. The difference between the detection rates is smaller for 16 test steps (5.9%) than for 91 test steps (12.2%). According to Figure 2, this is to be expected because the difference between the critical test values in 16 test steps is lower than in 91 test steps. The difference between the detection rates for the H_02 condition and for the Bonferroni correction presented in Table 2 is highly significant ($p < 0{,}0001$, Fisher's exact test) for 91 test steps and 16 test steps as well. These results confirm the expectation that a higher probability of a true detection of the alternative hypothesis is achieved with the critical test values determined by means of the method presented in this study, in comparison to the test values corrected in accordance with Bonferroni. This is also illustrated by the ROC-curves in Figure 4, which shows the ranking of the performance of the three test conditions described above. The ROC-curve for the single testing of the sample of 100 epochs (H_01) runs at the highest level indicating the best performance, whereas the ROC-curve for the Bonferroni correction (H_01_{Bonf91}) runs below the other curves and thus reveals the poorest performance. A critical test value determined by means of the method presented in this study (H_02_{91}) leads to a performance lower than that of the single testing but higher than that reached by the Bonferroni correction.

Because the time required for the detection of a response (number of epochs required) is closely related to the higher probability of a true detection of the alternative hypothesis, the mean detection time using critical test values corrected in accordance with Bonferroni as stated in Table 3 is also slightly longer than for those corresponding to the critical test values determined from H_02_{16} and H_02_{91}. The reduced examination time that results when using the more favourable critical test values is, however, not as important as the increased sensitivity achieved.

General discussion

The considerations and results demonstrated in this study for the modified Rayleigh test apply in principle to all statistical tests carried out in the time or the spectral domain used for the response detection in ASSR, ABR, or OAE. The modified Rayleigh test is one of the most effective tests among the one-sample tests used for response detection in the spectral domain (Cebulla et al, 2001). The critical test value corresponding to the number of test steps intended by the user can easily be found. In this study the steps of 16 and 91 have been used as examples. If another statistical test is used and the method to determine the critical test values for dependent samples described here is applied, only the simulations will have to be carried out for the construction of the corresponding H_02_x null-hypothesis described above. Here, x corresponds to the maximum number of test steps preferred by the user. The critical test value for the predefined error probability α can be determined directly from H_02_x.

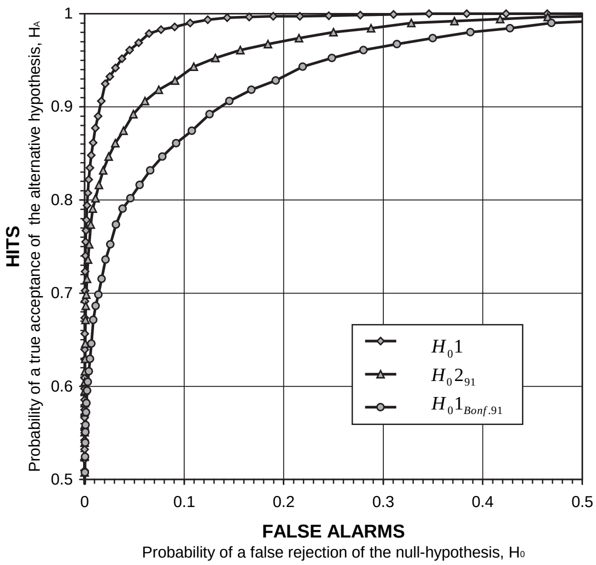

Figure 4. ROC-curves for the modified Rayleigh test applied under three different test conditions: (1) single testing using 100 epochs (H_01), (2) with the Bonferroni correction using 91 test steps ($H_01_{Bonf.\,91}$) and (3) using 91 test steps (H_02_{91}). Concerning the construction of the ROC-curves see text.

Both the error probability and the maximum number of test steps will depend on the actual application. The error probability $\alpha = 1\%$, which is used throughout this paper, is sufficient for the objective determination of the hearing threshold (e.g. by means of AMFR). A false rejection of the null-hypothesis is thus expected in one out of 100 recordings below threshold where there is no response. When determining the hearing threshold, such an incorrect measurement is not particularly critical. However, a considerably lower probability of a false rejection of the null-hypothesis is necessary for the hearing screening in newborns. With an error probability $\alpha = 0.1\%$, the hearing loss would still not be detected in one out of 1000 screenings on newborns with hearing impairment, but the consequences of a false pass would be dramatic for this one baby. Therefore, tests in hearing screening should preferably be carried out with an error probability even lower than $\alpha = 0.1\%$.

A critical test value, which has been corrected accordingly, has to be used to actually meet the required error probability, α, because the devices used in hearing screening usually work with a sequential test strategy. Every correction means an increase in the critical test value and, as a consequence, results in an increase in the probability of a false rejection of the alternative hypothesis. In hearing screening, this is synonymous with an increased fail rate, since a true response is not detected. The application of the familiar correction in accordance with Bonferroni does guarantee the required error probability, α, but the fail rate is increased excessively as a consequence of the overcorrection in the case of dependent samples. In contrast, the new method suggested here produces statistically-correct critical test values for dependent samples, as has been shown in this study. This procedure also leads inevitably to a slightly higher fail rate, but the probability of a false rejection of the alternative hypothesis is considerably lower than in the case of the correction in accordance with Bonferroni. The test parameters should be selected as effectively as possible in hearing screening, as well in objective determination of the hearing threshold, to compensate for the inevitable decrease in the probability of a true acceptance of the alternative hypothesis in a sequential test strategy.

If a maximum measuring time of two minutes is assumed in an ASSR hearing screening, the 120 seconds can either be divided into a larger number of short epochs or into a lower number of longer epochs. In an earlier article (Cebulla et al, 2001), we could show for the AMFR data which also has been used in the present study, that a higher number of short epochs is more effective in the case of one-sample tests. The detection rate shows a flat optimum at an epoch length of between one and two seconds. Apart from this, an epoch length that clearly exceeds 2 s extends the detection time. A response that is detected, for example, after 26 s at an epoch length of 1 s can only be detected after 30 s at an epoch length of 5 s, even in the most favourable case. Here, for example, the fact that every statistical test needs a minimum sample size for a reliable result has not been taken into consideration. Using the modified Rayleigh test, the first test of the sequential test procedure was carried out here with 10 epochs. With an epoch length of 5 s, this would mean that the first test has to await 50 s before it can be carried out. This would lead to a considerable lengthening of response detection times.

A stimulation level between 35 or 40 dB HL is used in ABR hearing screening. These stimuli are at supra-threshold levels for babies with normal hearing. Response detection is normally achieved within a relatively short measurement period. For example, the mean detection time is approximately 40 s (Cebulla et al, 2003; lecture presented at the IERASG meeting, Tenerife, 2003) with the ASSR algorithm (Stürzebecher et al, 2003) implemented on the MAICO, MB11 automated test system equipped with the BERAphone®. Here, a step-width of 1 is an obvious choice in the case of a sequential testing (epoch length about 1 s). The conditions that have to be considered during the objective determination of the hearing threshold by means of AMFR are different. Measurements above, close to and below the threshold have to be taken into account. Without a sequential test strategy, the examination would require an unreasonably high expenditure of time because a defined sample size would then have to be selected to enable responses close to the threshold to be detected. As the detection times for supra-threshold responses are relatively small (see above), this results as a whole in a sufficiently short duration of the threshold assessment when using a sequential test strategy, particularly because several frequencies can be tested simultaneously. According to our experience, a maximum test length of about six minutes seems to be sufficient before the test is discontinued with the result 'no detectable response'. With a measuring length of 6 min, an epoch length of 2 s and a test step-width of 2 should be chosen so that the number of sequential test steps is not too high. With these parameters, the time-effectiveness of the sequential test strategy becomes only slightly unfavourable. Picton et al (2003) reported that tests aimed at detecting AMFR close to the threshold had to be extended by up to 10 min or more. Dimitrijevic et al (2002) worked with periods up to 17 min. A greater step-width would be recommendable with these test durations. This, however, produces unnecessarily long test durations in supra-threshold responses. Furthermore, a test length of 10 min and more leads to a not justifiable total examination time. It is more favourable to use a statistical test procedure with a higher test power instead of lengthening the measuring time.

Until now, information from the higher harmonics is neglected, and one-sample tests that only evaluate the first harmonic have most often been used to detect ASSR in the frequency domain (Champlin, 1992; Dobie & Wilson, 1993; Lins & Picton, 1995; Aoyagi et al, 1999; John and Picton, 2000; Rance & Rickards, 2002). If, however, a q-sample test which also incorporates higher harmonics is used, a higher detection rate and a shorter detection time are achieved in comparison to the one-sample tests (Cebulla et al, unpublished). As recommended above, a maximum test duration of about six minutes can then be expected.

References

Aoyagi, M., Suzuki, Y., Yokota, M., Furuse, H., Watanabe, T. & Ito, T. 1999. Reliability of 80-Hz amplitude-modulation following response detected by phase coherence. *Audiol Neurootol*, 4, 28–37.

Cebulla, M., Stürzebecher, E. & Wernecke, K.D. 2000. Objective detection of auditory brainstem potentials. *Scand Audiol*, 29, 44–51.

Cebulla, M., Stürzebecher, E. & Wernecke, K.D. 2001. Objective detection of the amplitude modulation following response (AMFR). *Audiology*, 40, 245–252.

Cebulla, M., Stürzebecher, E., Berger, E.R., Shehata-Dieler, W. & Neumann, K. 2003. A new, rapid ABR screening algorithm based on click-evoked ASSR. XVIII IERASG Biennial Symposium, Puerto de la Cruz, Teneriffa.

Champlin, C.A. 1992. Methods for detecting auditory steady-state potentials recorded from humans. *Hear Res*, 58, 63–69.

Dimitrijevic, A., John, M.S., Van Roon, P., Purcell, D.W., Adamonis, J., et al. 2002. Estimating the audiogram using multiple auditory steady-state responses. *J Am Acad Audiol*, 13, 205–224.

Dobie, R.A. & Wilson, M.J. 1993. Objective response detection in the frequency domain. *Electroenceph Clin Neurophysiol*, 88, 516–524.

Don, M. & Elberling, C. 1996. Use of quantitative measures of auditory brain-stem response peak amplitude and residual back ground noise in the decision to stop averaging. *J Acoust Soc Am*, 99, 491–499.

Green, D.M. & Swets, J.A. 1966. *Signal Detection Theory and Psychophysics*. New York: John Wiley & Sons.

Hochberg, Y. & Tamhane, A.C. 1987. *Multiple Comparison Procedures*. New York: Wiley and Sons.

Howell, D.C. 1987. *Statistical Methods for Psychology*. Boston, MA: Duxbury Press.

John, M.S. & Picton, T.W. 2000. MASTER: a Windows program for recording multiple auditory steady-state responses. *Computer Methods and Programs in Biomedicine*, 61, 125–150.

Lins, G.L. & Picton, T.W. 1995. Auditory steady-state responses to multiple simultaneous stimuli. *Electroenceph Clin Neurophysiol*, 96, 420–432.

Lütkenhöner, B. 1991. Theoretical considerations on the detection of evoked responses by means of the Rayleigh test. *Acta Otolaryngol (Stockh)*, Suppl 491, 52–60.

Moore, B.R. 1980. A modification of the Rayleigh test for vector data. *Biometrika*, 67, 175–180.

Picton, T.W., John, M.S. & Dimitrijevic, A. 2002. Possible roles for the auditory steady-state responses in identification, evaluation and management of hearing loss in infancy. *Audiology Today*, 14, 29–34.

Picton, T.W., John, M.S., Dimitrijevic, A. & Purcell, D. 2003. Human auditory steady-state responses. *Int J Audiol*, 42, 177–219.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. 1986. *Numerical recipes: the art of scientific computing*. Cambridge: Cambridge University Press.

Rance, G. & Rickards, F. 2002. Prediction of hearing threshold in infants using auditory steady-state evoked potentials. *J Am Acad Audiol*, 13, 236–245.

Stürzebecher, E., Cebulla, M. & Wernecke, K.D. 1999. Objective response detection in the frequency domain-Comparison of several q-sample tests. *Audiology & Neurootology*, 4, 2–11.

Stürzebecher, E. 2001. *Method for Hearing Screening of Newborn by Means of Steady-State Responses Evoked with High Click Rate*. Patent Application 01610060.4 at the European Patent Office.

Stürzebecher, E., Cebulla, M. & Neumann, K. 2003. Click-evoked ABR at high stimulus repetition rates for neonatal hearing screening. *Int J Audiol*, 42, 59–70.

Valdes, J.L., Perez-Abalo, M.C., Martin, V., Savio, G., Sierra, C., et al. 1997. Comparison of statistical indicators for the automatic detection of 80 Hz auditory steady state responses. *Ear Hear*, 18, 420–429.