Doc Ophthalmol (2008) 117:163–170
DOI 10.1007/s10633-008-9121-1

ORIGINAL RESEARCH ARTICLE

A comparison of signal detection techniques in the multifocal electroretinogram

Tom Wright · Josefin Nilsson · Christina Gerth · Carol Westall

Received: 12 November 2007 / Accepted: 19 February 2008 / Published online: 7 March 2008

Abstract A common task in the analysis of the multifocal electroretinogram (mfERG) is determining which retinal areas have preserved signal in recordings which are attenuated by the effects of disease. Several automated methods have been proposed for signal detection from multifocal recordings, but no systematic study has been published comparing the performance of each. This article compares the sensitivity and specificity of expert human scoring with three different automated methods of mfERG signal detection. Recordings from control subjects were artificially modified to simulate decrease in signal amplitudes (attenuation) as well as total signal loss. Human scorers were able to identify areas with preserved signal at both low and high attenuation levels with a high specificity (minimum 0.99), sensitivities ranged from 0.2 to 0.94. Automated methods based on template correlation performed better than chance at all attenuation levels, with a slide fit method having the best performance. Signal detection based on signal to noise ratio performed poorly. In conclusion automated methods of signal detection can be used to increase signal detection sensitivity in the mfERG.

Keywords mfERG · Signal detection

Introduction

The multifocal Electroretinogram (mfERG) initially introduced by Sutter and Tran [1] is a sensitive tool for detecting and monitoring a wide range of retinal abnormalities [2]. Changes in the timing and amplitude of the biphasic waveform obtained, using the first order analysis of a simple stimulus, provides information relating to the function of different retinal layers [3]. Identifying and scoring the recorded waveforms from subjects with normal retinal function (control) is a relatively simple matter as the distinctive shape of the recording is easily recognised. Signal identification problems arise, however, when recording severely attenuated signals such as those produced by a disease compromised retina [4]. Previous studies have used several criteria for signal identification including minimum amplitude [5], and correlation with 'ideal' templates [6]. Comparison of different signal detection techniques in a population with retinal disease is difficult since it

T. Wright (✉) · J. Nilsson · C. Gerth · C. Westall
Department of Ophthalmology and Vision Sciences,
The Hospital for Sick Children, 555 University Avenue,
Toronto, ON, Canada M5G 1X8
e-mail: thomas.wright@sickkids.ca

T. Wright · J. Nilsson
Department of Clinical Neurophysiology, Sahlgrenska
Academy, Göteborg University, Gothenburg, Sweden

C. Westall
Department of Ophthalmology and Vision Sciences,
University of Toronto, Toronto, Canada

is hard to accurately quantify localised retinal function. In order to compare the performance of three different scoring methods with that obtained by expert human scorers this article uses data from control subjects with normal retinal function which is modified in a predictable manner to simulate disease processes.

Methods

The data sets

Data from 28 control subjects were collected on a VERIS^{TM} (EDI, San Mateo, CA) mfERG system. All procedures were approved by the research ethics board at the Hospital For Sick Children and the study conformed to the tenants of the Declaration of Helsinki. After obtaining informed consent data was collected according to ISCEV recommendations [7], using a standard scaled 103 hexagon stimulus presented at 75 Hz M-sequence exponent = 15. Luminance of white hexagons was approximately 200 cd.m^2 and dark hexagons 0 cd.m^2 resulting in an average luminance of 100 cd.m^2. Prior to testing pupils were dilated to at least 8 mm diameter using 1% Mydriacyl. mfERGs were recorded using a bipolar Burian-Allen electrode. Data were amplified by 50000 times and bandpass analog filtered (10–300 Hz). Two passes of the iterative artefact rejection protocol implemented in the Veris^{TM} software were applied with no further spatial averaging. This allowed 200 ms traces to be extracted from each hexagon before cross contamination between hexagons becomes an issue. All recordings were exported for post processing using custom scripts written in R [8]. Each recording was processed to simulate loss of retinal function. Previous work has shown that the retinal signal occurs within the first 60 ms after stimulation [9, 10]. Thus, we divided each trace recording into two epochs, a signal epoch 0–80 ms and a noise epoch 110–190 ms. Recordings were processed to replace the signal epoch in a random number of hexagons leaving 5–20 hexagons with signal. Then the amplitudes of traces from all hexagons were attenuated for a range of factors from 0 (no attenuation) to 5 (1/5th attenuation) (Fig. 1). Each of these processes is explained in further detail below.

Creation of noise hexagons

In order to simulate localised retinal function loss between 83 and 98 hexagons were randomly selected from each recording to have the signal removed. For each of these traces the signal epoch (0–80 ms) was replaced with a noise epoch (110–190 ms) from a hexagon of similar eccentricity.

Signal attenuation

In order to reduce the signal to noise ratio (SNR) of the recordings to simulate global retinal degeneration the noise epoch for each trace was multiplied by a factor and added to the signal epoch. The resultant epoch was then divided in order to obtain a waveform of similar amplitude to the original with the contribution of the signal reduced. The factor used to normalise the waveforms was dependent on how much noise was added. If the resultant waveform consisted of 1× signal epoch + 3× noise epoch this was divided by 4 to result in approximately the same root mean square (RMS) as the original wave. The resulting epoch was then used to replace the signal epoch in the waveform. It is worth noting that as the signal epoch will already contain some noise, the additive nature of random noise means that the factors used here can only be an estimate of the true proportion of noise in the final trace. The actual proportion of noise will tend to be slightly smaller that that indicated by the factors used in this text.

Signal detection methods

Human scoring

Recordings attenuated with factors of 0 and 2 were presented in a web interface (56 recordings in total). Four experienced electrophysiologists (5–20 years, mean 12.5 years experience) were asked to identify which hexagons in a recording still contained a signal. Participants were not given any extra information about the recordings such as what level of attenuation had been applied or how many hexagons still contained signal. The presentation mode was designed to closely resemble the 'traces' presentation of the Veris^{TM} software. As well as presenting all 103 hexagon traces at one time, each individual trace could be examined in more detail by holding a cursor

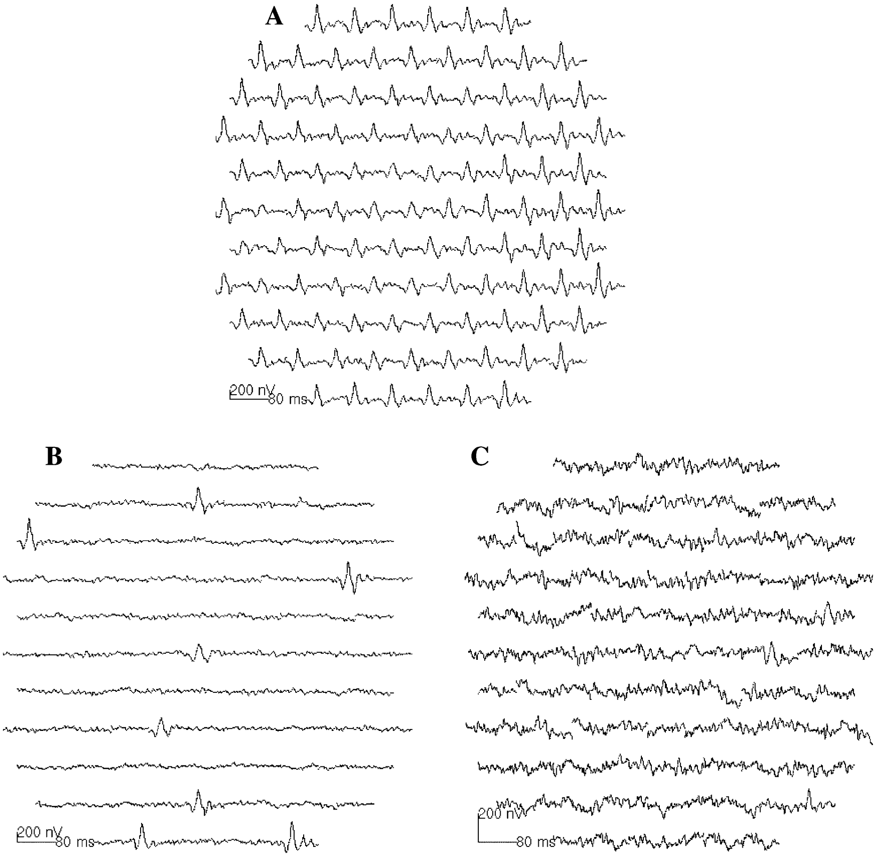

Fig. 1 Post processing of an individual mfERG recording. (**a**) The raw mfERG recording from one subject. (**b**) Recording after removal of signal from 95 hexagons. (**c**) The same recording after attenuation by a factor of 5 (note the difference in amplitude scales)

over the trace. The observer was requested to click on traces with a signal. Identified hexagons were surrounded with a border and could be deselected by clicking. Recordings were presented in a masked random order.

Automated scoring

Three automated scoring methods were studied (described below), the first two rely on comparing a template waveform to the traces recorded from each hexagon, and the third calculates an SNR value for each hexagon.

(1) *Waveform sliding* (*additive scaling*). This is a two-stage process. Initially a template waveform is generated; ISCEV guidelines suggest the template should be generated from age similar control data [7]. We used a template generated by averaging all hexagons from all 28 original recordings. A window from this waveform containing the area of interest is selected (6–46ms), which is then compared at incremental time windows with the trace to be scored (the target). The time with the least difference as measured by RMS difference between the waveforms is considered as the optimal time (Fig. 2). The template window is then stretched in amplitude to obtain the best-fit. As all the data being examined are based on control recordings, we do not expect a large time difference. thus the search window was set between −5 and +5 ms with steps of 0.1 ms for the implicit times. Amplitudes were expected to be more variable due to the attenuation process. Therefore, the best-fit amplitude was searched for with factors between 0.1 and 10 with a step size of 0.1. The final RMS difference between the template and target trace was used as a measure of the likelihood of the target trace containing a signal. For example, a RMS difference of 0 indicates a perfect fit between the template and the target trace. The waveform would be considered to contain a definite signal. Alternatively, a very high RMS difference would indicate the template waveform could not be modified to give a good fit between the recorded wave and the template, indicating no signal was present in the target.

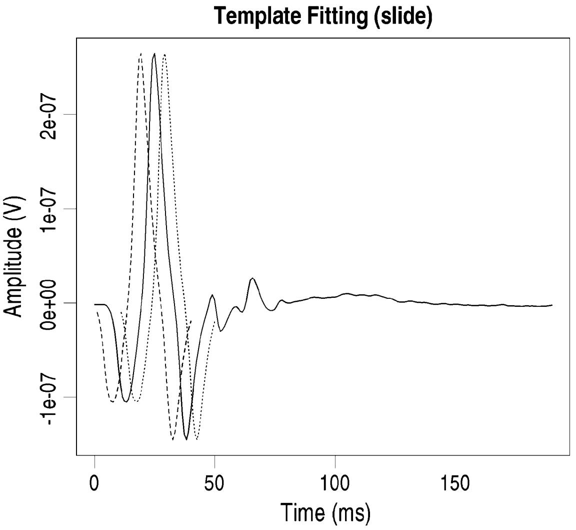

Fig. 2 Slide fitting of a template to the mfERG trace. The dashed line shows the template window of 6–43 ms slid by −5 ms and the dotted line shows the same template window slid by +5 ms

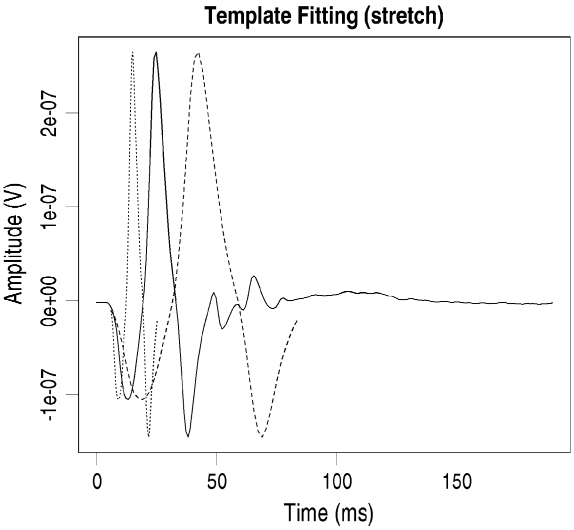

Fig. 3 Stretch fitting of a template to the mfERG trace. The dashed line shows the template window of 6–43 ms stretch by a factor of 0.5 and the dotted line the same window stretched by a factor of 2

(2) *Waveform stretching* (*multiplicative scaling*). This method is a modification of that described by Hood and Li [6]. Instead of the template window being incrementally slid along the target waveform to determine the best-fit for time, as described previously, the start point of the template is fixed and only the endpoint is moved. This leads to a stretching of the template with later components being moved further in time than earlier components. Again a window of 6–46 ms was used and this was stretched by a range of factors from 0.5 to 2 in steps of 0.01 (Fig. 3) to obtain the best temporal fit. The waveform is then stretched in amplitude as before. As with the slide method the final RMS difference between the template and target trace represented the likelihood of the target trace containing a signal.

(3) *Signal to noise ratio.* As stated earlier the recording settings used allowed a 190 ms trace to be extracted from each hexagon. An epoch of 0–80 ms was considered as containing the signal. An epoch of 110–190 ms was considered as containing only random noise (Fig. 4). The RMS for each of these epochs was calculated as a measure of the magnitude for each epoch. The ratio $\mathrm{RMS}_{\mathrm{signal}}/\mathrm{RMS}_{\mathrm{noise}}$ was calculated. If a waveform contained no signal the SNR would be close to 1. Higher SNR values would indicate an increased likelihood that the waveform contains a signal.

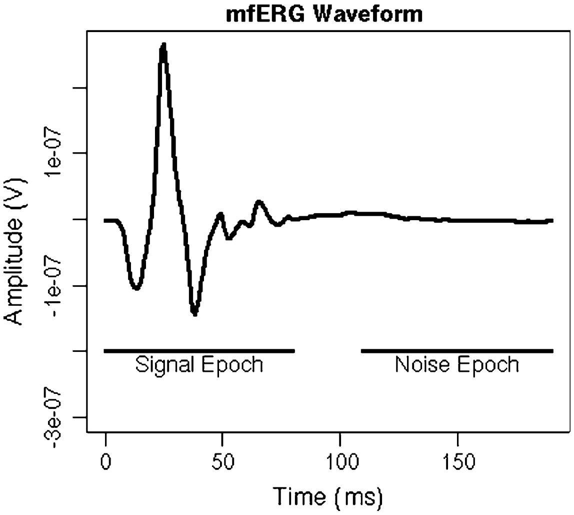

Fig. 4 Signal and noise epochs. Signal epoch is defined as the first 80 ms of the trace and the noise epoch as the last 80 ms

Analysis

Each of the continuous factors i.e. fit values from sliding and stretch fit methods and the SNR values, were scaled for each set of 28 recordings at a particular attenuation level to give a range of values between 0 and 1. These values were then compared with the truth values i.e. which hexagons actually contained the signal component, to determine the sensitivity and specificity. These were then plotted as receiver operator curves (ROC) and the area under the curve calculated to give a measure of performance. The sensitivity and specificity was also calculated for the human scorers at each attenuation level. The discrete nature of the human scoring (i.e. hexagons either do or do not contain signal) coupled with the small number of scorers makes drawing a smooth ROC curve impossible, therefore, this data is presented as individual points.

Results

Not surprisingly the performance of all the methods decreased as the attenuation was increased. The template scoring methods performed better than would be expected if scoring were performed purely by chance, even when the original signals were attenuated to 1/5th of their original size. The SNR method showed improvement to an attenuation factor of 2 (Fig. 5).

The slide method of template fitting showed the best performance over all attenuation levels with SNR fitting performing consistently worst (Table 1).

Human scoring of the recordings at 0 attenuation and at an attenuation factor of 2 showed a very high specificity i.e. a low number of false positives; however the sensitivity showed a large range (Table 2). The sensitivity levels achieved by the automated scoring methods at the 0.99 specificity level are also shown for comparison (Table 3).

Figure 6 shows the same information plotted as datapoints on ROC graphs for the slide fit method. It can be seen that human scorers only outperformed the automated method once (blue circle) and only with data that were not attenuated.

Intra-class correlation shows significant agreement between human scorers at both tested attenuation levels 0.71 ($P < 0.01$) at 0 attenuation and 0.61 ($P < 0.01$) at the 0.5 attenuation level (two way model, test for agreement) [13].

In order to assess the degree of agreement between the automated scoring methods and human scoring, a

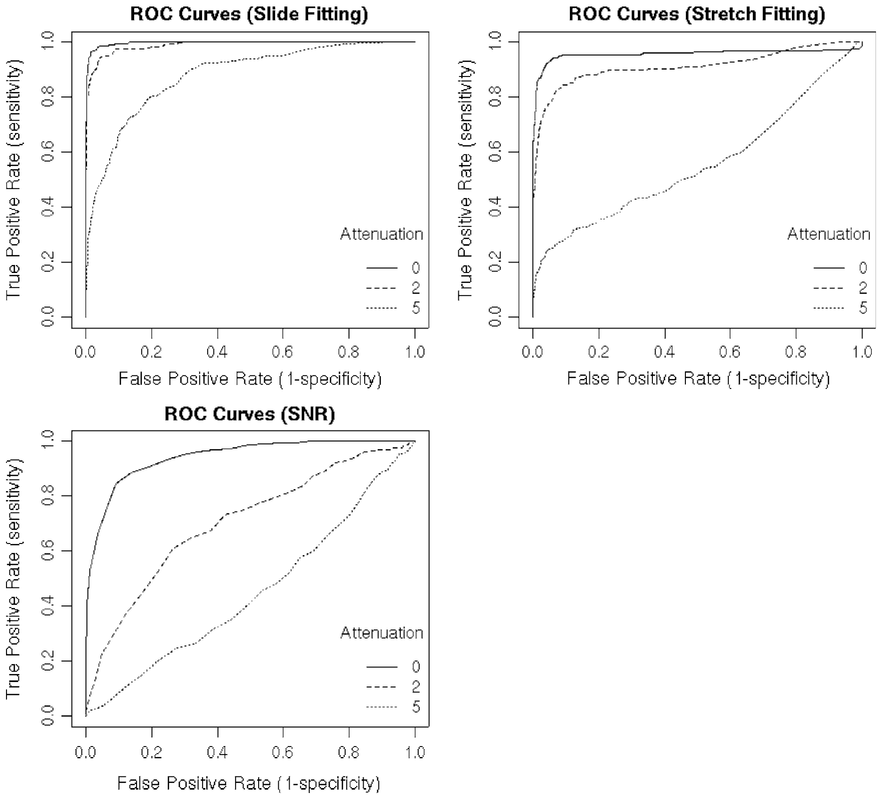

Fig. 5 Receiver operator curves. ROC curves indicate the sensitivity and specificity for each automatic scoring method

Table 1 Area under the curve values for each of the automated scoring methods, at each attenuation level

	Attenuation		
Method	0	2	5
Slide	1	0.99	0.88
Stretch	0.96*	0.91*	0.55*
SNR	0.94*	0.70*	0.45*

* Indicates siginifcant differences as compared with the slide method ($P < 0.01$) [11, 12]

Table 2 sensitivity and specificity at each attenuation level for 4 human scorers

Scorer	Attenuation			
	0		2	
	Sensitivity	Specificity	Sensitivity	Specificity
1	0.37	1	0.2	1
2	0.94	0.99	0.74	1
3	0.68	0.99	0.48	0.99
4	0.69	1	0.49	1

Table 3 Sensitivity for automated scoring methods at 0.99 specificity level

	Attenuation	
Method	0	2
Slide	0.95	0.86
Stretch	0.66	0.60
SNR	0.53	0.06

* Indicates that value is significantly different from sensitivity obtained using the slide method ($P < 0.01$)

cut off value was selected for each automatic scoring method at each attenuation level to obtain a specificity of 0.99 (the lowest obtained by any human scorer). These values were then used to identify which hexagons were considered to contain signal. The results were then compared to human scorer 2 (the scorer obtaining the highest sensitivity and specificity) and intra-class correlations for agreement calculated (Table 4).

Discussion

We were able to calculate sensitivity and specificity for three automated scoring methods as well as for human expert scorers, by artificially modifying mfERG recordings from control subjects, enabling knowledge of which hexagons truly contained signals. Numbers for sensitivity and specificity may be considered counterintuitive in this study, since positive signal detection indicates a functioning retinal area, as opposed to community based screening where positive findings usually imply disease. Here, a low sensitivity represents functioning retinal areas not being detected. In a clinical situation this will lead to over diagnosis of retinal dysfunction. Conversely a test with a low specificity will identify retinal areas as producing a signal where none is present; this will probably lead to under diagnosis.

Inter-class correlations show a high level of agreement between the human observers. The very

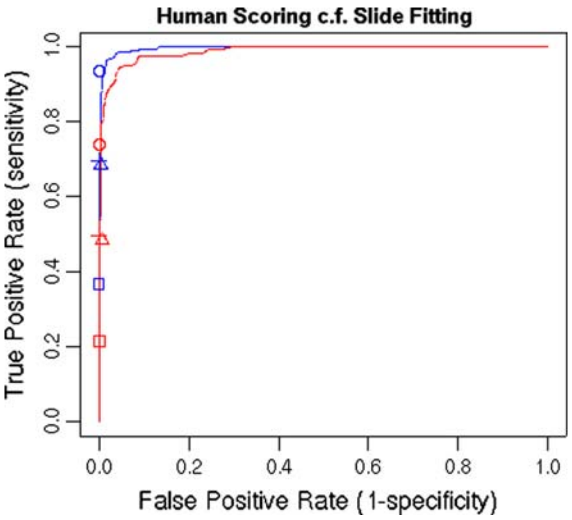

Fig. 6 Human sensitivity and specificity compared with sensitivity and specificity obtained by slide fitting. Blue data indicates results from unattenuated data, red data results from data attenuated with a factor of two. Scorer 1 is represented by squares, Scorer 2 circles, scorer 3 triangles and scorer 4 a + symbol

Table 4 Inter-class correlations (two way model) showing agreement between automated scoring methods and human scorer 2

	Attenuation	
Method	0	2
Stretch	0.80 ($P < 0.01$)	0.69 ($P < 0.01$)
Slide	0.89 ($P < 0.01$)	0.79 ($P < 0.01$)
SNR	0.02 ($P = 0.41$)	0.02 ($P > 1$)

high specificity levels indicate that trained observers rarely confuse noise with signal. The large variation in sensitivity is hard to explain but may reflect the degree of conservatism in signal selection exhibited by the observers. It is worth noting that scorer 2 who had the best performance in terms of sensitivity and specificity had a background in clinical neurophysiology rather than a visual electrophysiology, as such, they may be more practiced in identifying very small signal components in recordings. More scorers would be useful in elucidating the nature of the variation.

Interestingly, the use of SNRs as a scoring method deteriorated very rapidly as attenuation levels increased. This method suffers in comparison with the template correlation methods from having no prior knowledge of expected wave morphology.

With the dataset used in this study, the slide fitting method performed best. This may be explained by the fact that all the data are based on recordings from normal retinas and no changes in the wave morphology have been introduced. It has been suggested that the template stretching method gives a better fit to the waveform in certain disease processes such as diabetes [14] and retinitis pigmentosa [10], this study does not invalidate those results and it is probably the case that prior knowledge of the expected effects of a disease process can direct the selection of signal detection method to increase sensitivity further.

The very poor agreement between human scorers and the SNR method are due to the very low sensitivity of the SNR method at the 0.99 specificity level. Earlier work in patients with retinitis pigmentosa suggests that the SNR methods may have a role in very attenuated recordings where traces are too severely attenuated for template correlation methods to perform reliably. In this previous study, hexagons containing signal were identified by a SNR based scoring method. When hexagons identified by this method were averaged together, the typical mfERG waveform morphology became apparent, averaging other hexagons did not give produce a recognisable waveform [4].

While the presentation methods used for the human scoring task were designed to resemble the traces presentation of the Veris software, it was not possible to implement all the tools available in this software, such as direct comparison with control data, which the human observer can use to increase their sensitivity. Conversely human observers were aware that all the recordings presented contained some traces with signal and so may have been tempted to identify signals in recordings which, under other circumstances, may have simply been described as unscoreable. The methods implemented to simulate retinal dysfunction could only modify the relative amplitude of the signal epoch compared with the noise epoch. While this is a good model for many diseases of the retina, it does not investigate the effect of diseases such as diabetes which can change waveform morphology.

In conclusion, template based methods of signal detection have a role to play in increasing the sensitivity of the mfERG test especially in situations where the retinal signal is attenuated globally. Prior knowledge of the expected waveform changes can be useful in deciding exactly which method will give the best results.

References

1. Sutter EE, Tran D (1992) The field topography of ERG components in Man-I. The photopic luminance response. Vision Res 32:433–446
2. Lai TYY, Chan WM, Lai RYK, Ngai JWS, Li H, Lam DSC (2007) The clinical applications of multifocal electroretinography: a systematic review. Surv Ophthalmol 52:61–96
3. Hood DC (2000) Assessing retinal function with the multifocal technique. Prog Retin Eye Res 19:607–646
4. Gerth C, Wright T, Heon E, Westall CA (2007) Assessment of central retinal function in patients with advanced retinitis pigmentosa. IOVS 48:1312–1318
5. Hood DC, Holopigian K, Greenstein V, Seiple W, Li J, Sutter EE, Carr RE (1998) Assessment of local retinal function in patients with retinitis pigmentosa using the multi-focal ERG technique. Vision Res 38:163–179
6. Hood DC, Li J (1997) A technique for measuring individual multifocal ERG records. Opt Soc Am, Trends Opt Photon 11:33–41
7. Hood DC, Bach M, Brigell M, Keating D, Kondo M, Lyons JS, Palmowski-Wolfe AM (2007) Title. International society for clinical electrophysiology of vision. http://www.iscev.org/standards/pdfs/mfERGSguidelines9–07r3.pdf
8. Computing RFfS, R: A language and environment for statistical computing, 2.4.1, 2006, http://www.R-project.org
9. Hood DC, Frishman LJ, Saszik S, Viswanathan S (2002) Retinal origins of the primate multifocal ERG: implications for the human response. Invest Ophthalmol Vis Sci 43:1673–1685
10. Greenstein VC, Holopigian K, Seiple W, Carr RE, Hood DC (2004) Atypical multifocal ERG responses in patients with diseases affecting the photoreceptors. Vision Res 44:2867–2874

11. Hanley JA, McNeil BJ (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843
12. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36
13. de Vet HC, Terwee CB, Knol DL, Bouter LM (2006) When to use agreement versus reliability measures. J Clin Epidemiol 59:1033–1039
14. Schneck ME, Bearse MA Jr, Ying H, Barez S, Jacobsen C, Adams AJ (2004) Comparison of mfERG waveform components and implicit time measurement techniques for detecting functional change in early diabetic eye disease. Doc Ophthalmol 108:223–230