Doc Ophthalmol (2009) 119:67–78
DOI 10.1007/s10633-009-9171-z

ORIGINAL RESEARCH ARTICLE

Reproducibility of the mfERG between instruments

Wendy W. Harrison · Marcus A. Bearse Jr. · Jason S. Ng · Shirin Barez · Marilyn E. Schneck · Anthony J. Adams

Received: 15 August 2008 / Accepted: 3 March 2009 / Published online: 26 March 2009

Abstract *Purpose* First, to examine both the reproducibility of the multifocal electroretinogram (mfERG) recorded on different versions of the same instrument, and the repeatability of the mfERG recorded on a single instrument using two different amplifiers. Second, to demonstrate a means by which multicenter and longitudinal studies that use more than one recording instrument can compare and combine data effectively. *Methods* Three different amplifiers and two mfERG setups, one using VERIS^{TM} 4.3 software (mfERG1) and another using VERIS^{TM} Pro 5.2 software (mfERG2), were evaluated. A total of 73 subjects with normal vision were tested in three groups. Group 1 ($n = 42$) was recorded using two amplifiers in parallel on mfERG1. Group 2 ($n = 52$) was recorded on mfERG2 using a single amplifier. Group 3 was a subgroup of 21 subjects from groups 1 and 2 that were tested sequentially on both instruments. A fourth group of 26 subjects with diabetes were also recorded using the two parallel amplifiers on mfERG1. P1 implicit times and N1-P1 amplitudes of the 103 local first order mfERGs were measured, and the differences between the instruments and amplifiers were evaluated as raw scores and *Z*-scores based on normative data. Measurements of individual responses and measurements averaged over the 103 responses were analyzed. *Results* Simultaneous recordings made on mfERG1 with the two different amplifiers showed differences in implicit times but similar amplitudes. There was a mean implicit time difference of 2.5 ms between the amplifiers but conversion to *Z*-scores improved their agreement. Recordings made on different days with the two instruments produced similar but more variable results, with amplitudes differing between them more than implicit times. For local response implicit times, the 95% confidence interval of the difference between instruments was approximately ±1 *Z*-score (±0.9 ms) in either direction. For local response amplitude, it was approximately ±1.6 *Z*-scores (±0.3 μV). *Conclusions* Different amplifiers can yield quite different mfERG P1 implicit times, even with identical band-pass settings. However, the reproducibility of mfERG *Z*-scores across recording instrumentation is relatively high. Comparison of data across systems and laboratories, necessary for multicenter or longitudinal investigations, is facilitated if raw data are converted into *Z*-scores based on normative data.

Keywords Multifocal electroretinogram · Reproducibility · VERIS^{TM} instruments · Repeatability

W. W. Harrison (✉) · M. A. Bearse Jr. · J. S. Ng · S. Barez · M. E. Schneck · A. J. Adams
University of California Berkeley School of Optometry, Vision Science Program, 360 Minor Hall, Berkeley, CA 94720-2020, USA
e-mail: wwh@berkeley.edu

J. S. Ng
Department of Basic and Visual Science, Southern California College of Optometry, 2575 Yorba Linda Blvd., Fullerton, CA 92831, USA

Introduction

The multifocal electroretinogram (mfERG) is a non-invasive objective technique that simultaneously measures retinal function at multiple retinal locations. It is used for the evaluation of retinal neuronal populations, as well as for the prediction and assessment of a wide variety of retinal diseases, including retinitis pigmentosa, diabetic retinopathy, and age-related macular degeneration [1–6]. The mfERG is also used for the evaluation of drug toxicity and surgical success [7–10], and its uses continue to expand in both the clinical and research arenas.

Although limited in number, previous studies have examined the repeatability of the mfERG and found it to be high with variations across systems [3, 11–15]. These studies have reported implicit time coefficients of variation (CV) as low as 3.1% (when achieving good repeatability was a goal of the study), and as high as 30.3% (when factors influencing variations in the mfERG were not fully controlled) [12, 15]. The CVs for amplitudes have been reported to range from 10.4% to 36% [13, 15]. Most studies have found that averaging over larger retinal areas reduces variability, and have consequently reported the CVs of rings of responses. Given all of the potential sources of variability that exist in an mfERG recording session, both intrinsic to the subjects and in the stimulus conditions and equipment, the high repeatability from these past studies is encouraging as long as the testing environment is controlled. ISCEV guidelines for clinical mfERG recording [16] are in place to help achieve uniformity in testing situations.

While the ISCEV guidelines specify that each clinic or laboratory establish their own norms, they do not address how clinics or laboratories could pool data for multicenter investigations. These may be necessary in the future to improve the statistical power of mfERG studies in the presence of relatively small samples. In addition, malfunction or aging of the mfERG equipment being used in a clinic or laboratory can require replacing components, causing inconsistency in the data being collected. This is especially important if follow-up data are to be interpreted or in longitudinal studies over a number of years. Scientists and clinicians are faced with the dilemma of replacing aging equipment while attempting to reduce inconsistencies in data collection and interpretation.

Reproducibility of the mfERG across instruments has not previously been examined. The purpose of this study is to evaluate the robustness and stability of the mfERG as it is recorded over both time and with different instrumentation (in the case of this study, different VERIS^{TM} instruments and amplifiers). Our results show that the reproducibility of the mfERG across recording instrumentation is quite high and that converting raw data into *Z*-scores based on normative data facilitates meaningful comparison of results across recording systems and different laboratories.

Methods

Systems and stimulus characteristics

Two visual evoked response imaging systems (VERIS^{TM}) (EDI, Redwood City, CA) were used to record first-order mfERGs. Both systems stimulated using luminance modulation of a 45°, 103-element hexagonal array scaled with eccentricity. The stimulus background, bright flashes, and dark elements were set to 100 cd/m^2, 200 cd/m^2, and <2 cd/m^2 (99% contrast), respectively. In addition, the ambient room lighting was between 80 and 100 cd/m^2 on the wall behind each instrument. Both systems had a 75 Hz frame rate monochrome CRT monitor display and ran a standard m-sequence (2^{15}–1) that lasted approximately 8 min. Each recording session was broken into 16 segments, approximately 30 s each, and the retinal signals were band-pass filtered at 10–100 Hz and sampled every 0.83 ms.

However, some features were different between the two recording setups (Table 1). Features unique to the first system (mfERG1) include that it runs VERIS^{TM}

Table 1 Differences between mfERG instruments

Characteristics	mfERG1	mfERG2
Veris software	VERIS^{TM} 4.3	VERIS^{TM} Pro 5.2
Amplifier model(s)	CP511 and P511	LT15
Amplifier setting	100,000	50,000
Monitor display and screen resolution	CRT 75 Hz	CRT 75 Hz
	1024 × 768 pixels	640 × 480 pixels

4.3 software and has a stimulus screen resolution of 1024 × 768. It also has two external Grass Telefactor (Astro-Med Inc®, West Warwick, RI) amplifiers. The first amplifier ("mfERG1 New Amp"; recording channel 1; Grass model CP511) was produced in 1996. The second amplifier ("mfERG1 Old Amp"; recording channel 2; Grass model P511) was manufactured in 1983. Both amplifiers on mfERG1 were set to amplify 100,000 times. Features unique to the second system (mfERG2) include that it runs VERIS™ Pro 5.2 software, has a stimulus screen resolution of 640 × 480, and has one computer-controlled Grass amplifier ("mfERG2 Amp"; Grass model 15LT), which was produced in 2006 and set to a gain of 50,000.

Comparison of the frequency response curves of the amplifiers, as specified by Grass, showed that the two newer amplifiers (mfERG1 New Amp and mfERG2 Amp) should have similar band-pass filtering characteristics but that the older amplifier (mfERG1 Old Amp) is slightly different. The difference between the amplifiers of mfERG1 was verified by inputting sine waves of varying frequencies but fixed amplitude and measuring the output amplitudes with the filters set at 10–100 Hz (Fig. 1). In addition, an artificial eye comprising a photodiode and an R–C circuit was run on both instruments and all three amplifiers to further characterize the implicit time differences inherent between them. The peak latencies of the first order "mfERGs" recorded from the artificial eye were consistently 2.5 ms shorter for the older (mfERG1, channel 2) amplifier than for the other two amplifiers.

Subjects and recordings

Seventy-three subjects with normal vision and 26 subjects with diabetes were included in this study. Patient demographic information is given in Table 2. The subjects were divided into four groups. Group 1 comprised 42 subjects with normal vision recorded simultaneously (in parallel) on both amplifiers of mfERG1. Group 2 comprised 52 subjects with normal vision recorded on mfERG2. Within group 2, 9 of these subjects returned for follow-up 1 year later to examine intra-instrument repeatability over time. Group 3 was a subset of the first two groups and consisted of 21 subjects who were run on both instruments within a two month period (mean = 0.94 ± 0.68 months). Group 4 was composed of 26 subjects with diabetes without

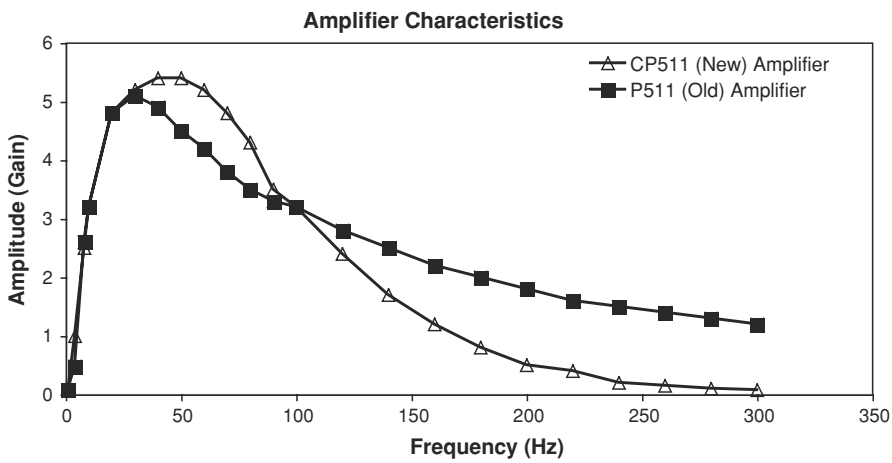

Fig. 1 The measured filtering characteristics of the new (Δ) and old (■) amplifiers on mfERG1 when set at 10–100 Hz

Table 2 Subject demographic information

Subject group	mfERG instrument	Number of subjects	Age ± SD
Group 1	mfERG1: both amplifiers	42 with normal vision	45.2 ± 12.75
Group 2	mfERG2	52 with normal vision	43.7 ± 14.5
Group 3 (subgroup of groups 1 and 2)	mfERG1 and mfERG2	21 with normal vision	47.4 ± 13.6
Group 4	mfERG1	26 with diabetes	51.3 ± 11.9

retinopathy, recorded on the two parallel channels of mfERG1.

MfERGs were recorded from one eye while the other eye was occluded. Pupils were fully dilated to at least 7 mm with 1% tropicamide and 2.5% phenylephrine, and 0.5% proparacaine was used to anesthetize the cornea prior to recording. A Burian-Allen bipolar contact lens electrode filled with 1.0% carboxymethlcellulose sodium solution was used for all mfERG recordings. Each instrument was used with its own dedicated contact lens electrode. A clip ground electrode was applied to the subject's earlobe and the resistance between the electrode leads was measured and kept under 10 k-Ohms. Both systems had in-line video cameras that allowed for real-time observation of the eye during testing. Recording segments contaminated by signal saturation or loss of fixation were discarded and repeated. All subjects had 20/20 (logMAR 0.0) or better visual acuity and were free of retinal disease and media opacities, as evaluated by ophthalmic examination and masked retinal photograph grading. All subjects had refractive errors between −6D and +4D. The study adhered to the Declaration of Helsinki and was approved by the Committee for the Protection of Human Subjects at the University of California Berkeley. Written informed consent was obtained from all subjects after the study was fully explained at their first visit.

Waveform and data analysis

The first-order mfERG kernel was analyzed. A single iteration of artifact removal was used on both instruments with 17% spatial averaging. The 103 mfERGs were exported and the Hood and Li template scaling method was applied to all waveforms to derive P1 implicit time and N1-P1 amplitude [17]. This method minimizes the least squares difference between a waveform and the local template. The template represents the mean local waveform of the subjects with normal vision and it is independently scaled in both amplitude and time to fit the individual local responses. The scaling factors are then used to derive implicit time and amplitude. The templates were created from the data of all subjects with normal vision in a group and a different set of 103 local response templates was used for each of the three amplifiers. Group 4's data was analyzed using the appropriate template from group 1. Implicit times and amplitudes were evaluated for all subjects as both raw scores and *Z*-scores, where the mean and standard deviation were calculated from all subjects with normal vision available for that amplifier-instrument combination after determining that the normative data for each of the 103 hexagons did not differ from a normal distribution (chi-square tests; mean $P = 0.58 \pm 0.25$). Responses were analyzed as whole eye averages (103 response measures averaged together) and also as individual local mfERG measurements.

Results

Amplifier comparisons on the same mfERG instrument

Recordings from the 2 amplifiers of mfERG1 using the two parallel channels were made from the subjects in groups 1 and 4. As the two recordings were made simultaneously, any differences between them can be attributed to the amplifiers (potential differences in gain, filtering, and noise), and no other sources of variation existed. The raw measurements of the 103 local mfERGs were first examined and then they were converted to *Z*-scores. The 103 raw measurements were then averaged to give one value for each subject. The *Z*-scores were similarly averaged.

The N1-P1 amplitudes were very similar with a mean difference of 0.01 ± 0.003 μV, and a maximum difference of 0.013 μV (6.5% of the mean value) between amplifiers for the whole eye average of any individual subject (data not shown). This was expected since the amplifiers were calibrated to provide similar overall gains. The mean amplitudes for the two channels were also similar for individual hexagons (0.21 ± 0.05 μV and 0.20 ± 0.05 μV for the first and second channels, respectively). Figure 2a shows the whole eye raw amplitude data obtained with both amplifiers from the subjects with normal vision (group1) and subjects with diabetes (group 4). Figure 2b shows the *Z*-scores of these same subjects. Figure 2b illustrates how the small difference between the amplifiers decreased after the conversion to *Z*-scores, and the data for both groups fall along a diagonal with a slope of 1.

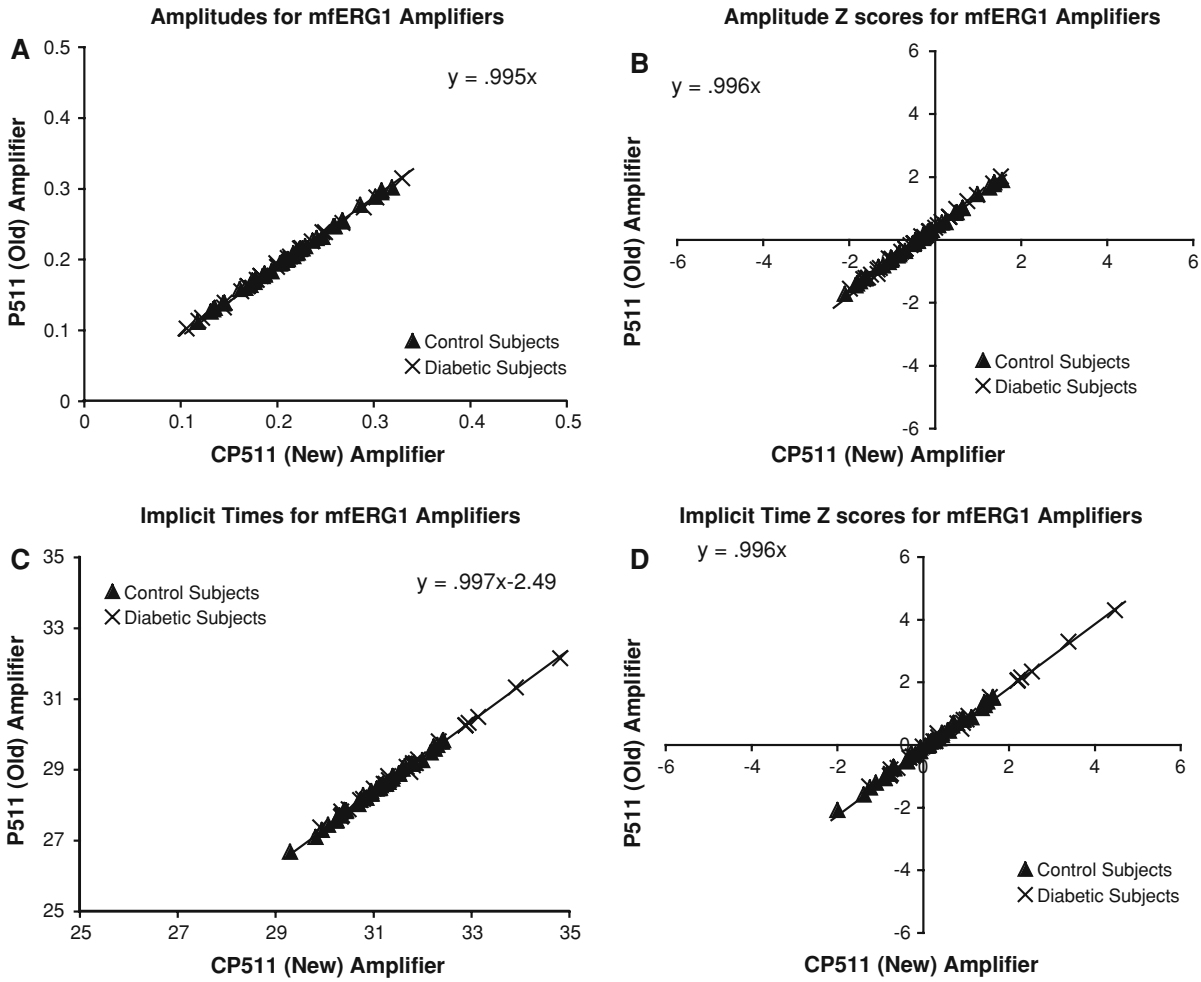

Fig. 2 mfERG1 amplifier raw data and *Z*-score comparison for amplitude and implicit time. **a** Raw amplitude data; **b** Amplitude *Z*-score data; **c** Raw Implicit Time data and **d** *Z*-score Implicit Time data. Each data point indicates a whole eye average for one subject. Subjects with normal vision, Group 1 (Control) (▲), and subjects with diabetes, Group 4 (X), are plotted together

Figure 2c shows the raw implicit time data obtained from both amplifiers. The mean implicit time difference between the two amplifiers was 2.5 ms. Figure 2d shows the *Z*-scores of the diabetic subjects and the subjects with normal vision with the data falling on a diagonal (slope = 1) passing through the origin. The implicit times showed a better agreement after the conversion to *Z*-scores.

Local response implicit time differences between the amplifiers were examined for subjects with diabetes. The 2.5 ms mean difference in implicit times between the two amplifiers also occurred locally, but with conversion to *Z*-scores, the amplifiers had good local agreement for all simultaneous recordings. Past studies in our lab have used implicit time *Z*-scores ≥ 2.0 ($P \leq 0.023$) as indications of abnormality [3, 18]. Table 3 shows that by applying

Table 3 Local amplifier agreement for subjects with diabetes (95.6%)

	New Amp >2 *Z*	New Amp <2 *Z*	Total
Old Amp >2 *Z*	401 (15.0%)	53 (2.0%)	454 (17%)
Old Amp <2 *Z*	63 (2.4%)	2161 (80.6%)	2224 (83%)
Total	464 (17.4%)	2214 (82.6%)	2678 (100%)

this criterion to the local data from the subjects with diabetes in this study, the two amplifiers had 95.6% agreement when classifying a local mfERG implicit time as normal or abnormal. A similar analysis was done for the subjects with normal vision using a criterion of 1.0 *Z*-score, also producing a high agreement of 92.5% (data not shown).

Reproducibility between different instruments

This section presents the comparison of mfERG data collected on different days using different instruments. Figure 3 shows the results for whole-eye average comparisons between the older amplifier (Grass model P511) of the mfERG1 system and mfERG2 system (Grass model 15LT amplifier) for subject group 3. The plot of the implicit times in Fig. 3a shows that, on average, there is a 2.5 ms difference between the two instruments, which is in agreement with the artificial eye. The mean implicit time of subjects on mfERG1 Old Amp was 28.80 ± 0.91 ms and the mean implicit time on mfERG2 was 31.30 ± 0.87 ms. As expected, there is a lower correlation ($R^2 = 0.81$) in the implicit time data than was observed earlier in the simultaneous recordings on a single mfERG instrument. The implicit times obtained on the two instruments are re-plotted as a Bland–Altman plot [19] in Fig. 3b. The difference between the two instruments is plotted on the *y*-axis and the mean of the instruments is plotted on the *x*-axis for each subject. The zero slope (95% CI = −0.24 to 0.20) and the *y*-intercept of the least squares regression indicate that there is an implicit time offset of about 2.5 ms between them. (If the intercept and the slope of the line were both 0, the two instruments would be directly comparable. If the line had a significant slope, the instruments would not be easily comparable.) By converting the implicit times into *Z*-scores, the two instruments are now more comparable (Fig. 3c). The 95% confidence interval of ±0.86 *Z*-scores indicates that implicit time *Z*-scores are highly reproducible on the two instruments. The differences between the two instruments ranged from 0.06 to 1.03 *Z*-scores.

Figure 4a shows the whole eye average amplitude comparison for the 21 subjects in group 3 for the older amplifier of mfERG1 and for mfERG2. The mean amplitude of mfERG2 was 0.30 ± 0.07 μV compared to 0.20 ± 0.05 μV for mfERG1 Old Amp.

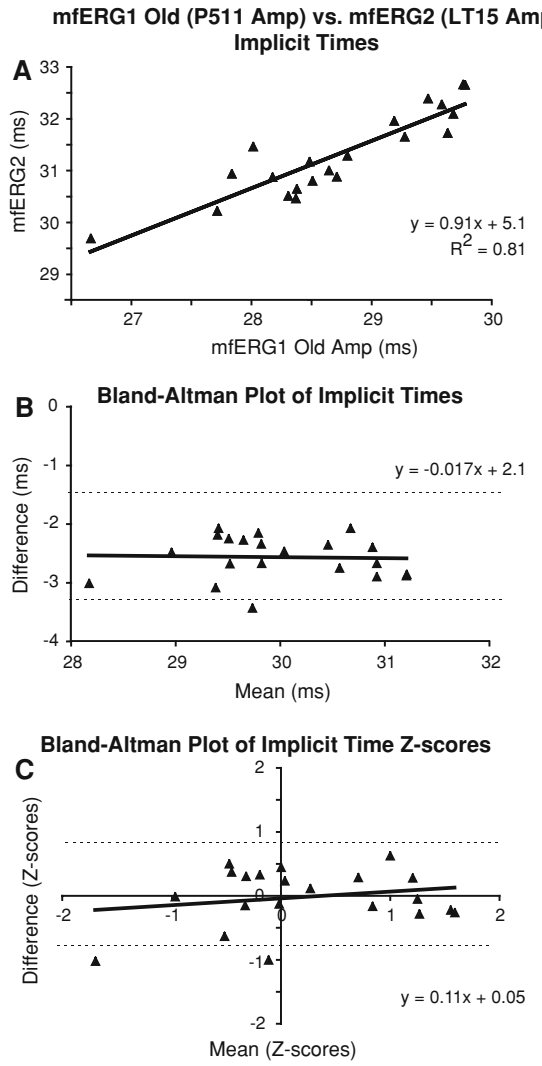

Fig. 3 Implicit time comparison for the older amplifier of mfERG1 and mfERG2. **a** Comparison of the implicit times (ms). Each point is a whole eye average of one subject from group 3; **b** A Bland–Altman plot of that same data as **a**. The dashed lines on the plot indicate the 95% confidence interval and the solid line is the mean difference (2.5 ms for the range of the mean implicit time data observed) between the instruments for all 21 subjects. The slope of the line is not statistically different from zero ($P = 0.88$); **c** The Bland–Altman plot of the *Z*-scores of the implicit time data with the dashed lines indicating the 95% confidence interval and the solid line indicating the mean difference (0.05 *Z*-score units) for the 21 subjects. The slope of this line is not different from zero ($P = 0.92$)

Although the agreement of amplitudes between the instruments varies among the subjects ($R^2 = 0.43$), the two instruments are comparable as the 95% confidence interval of the slope of the regression line

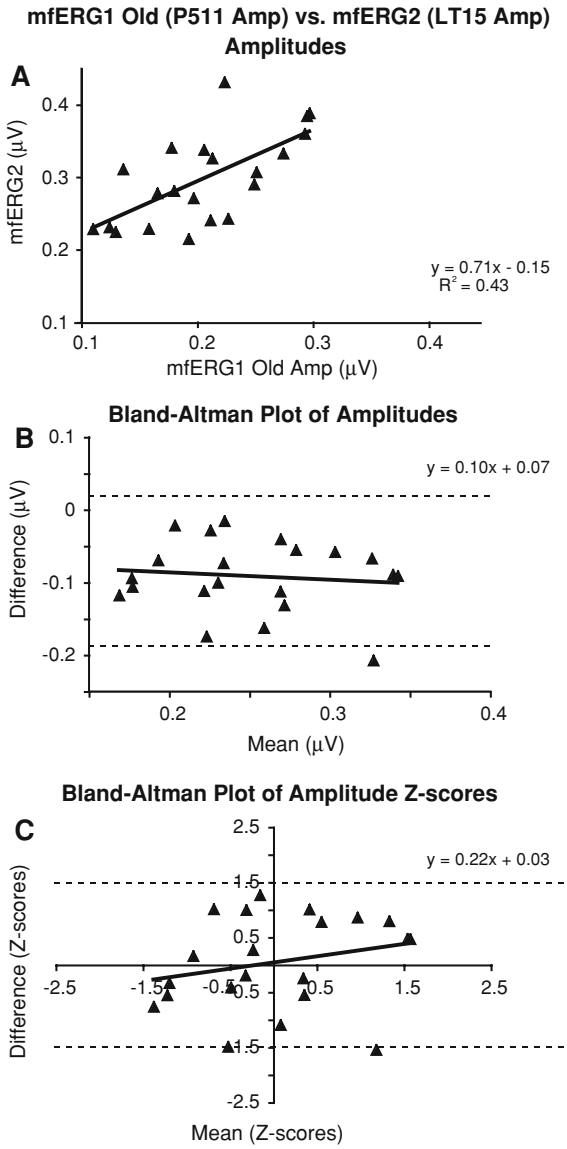

Fig. 4 Amplitude comparison for the older amplifier of mfERG1 and mfERG2. **a** Comparison of amplitudes (μV). Each point is a whole eye average of one subject from group 3. **b** A Bland–Altman plot of that same data as **a**. The dashed lines on the plot indicate the 95% confidence interval and the solid line is the mean difference (0.07 μV) between the instruments for all 21 subjects. The slope of this line is not statistically different than zero ($P = 0.23$). **c** The Bland–Altman plot of the *Z*-scores of the amplitude data with the dashed lines indicating the 95% confidence interval and the solid line indicating the mean difference (0.03 *Z*-score units) for the 21 subjects. The slope of this line is not different from zero ($P = 0.29$)

contains 1.0 (95% CI = 0.32–1.11). The Bland–Altman plot of the amplitude data (Fig. 4b) shows an average difference of 0.07 μV across all values but with a large 95% confidence interval associated with this value (0.01 to −0.19 μV). The *Z*-score Bland–Altman plot (Fig. 4c) also has a slope that is not significantly different from zero ($P = 0.29$) and the fact that the regression line passes through zero shows that the data is in better agreement with this conversion. The range of amplitude differences for these 21 subjects was large ranging from 0.1 to 1.6 *Z*-scores. The 95% confidence interval of ±1.5 *Z*-scores indicates that amplitude is not as reproducible as implicit time.

Data collected on the newer amplifier of mfERG1 (Grass model CP511) was also compared to data collected on mfERG2. As expected from their similar band-pass filtering characteristics, these two amplifiers exhibited raw implicit times that were similar (Fig. 5a), with a mean difference of only 0.1 ± 0.34 ms between the two instruments (Fig. 5b). The mean implicit time for mfERG1 New Amp was 31.40 ± 0.90 ms and the mean implicit time for mfERG2 was 31.30 ± 0.87 ms. Conversion of the data into implicit time *Z*-scores produced an even smaller mean difference between the instruments, making them more comparable (Fig. 5c). The 95% confidence interval of the difference between the two similar amplifiers was ±0.74 *Z*-scores.

The mfERG recordings performed to compare the instruments were not obtained in the same session, and so the question arises as to how much of the observed difference is due to subject variation over time and how much is due to actual instrumentation and electrode differences. To address this, 9 subjects with normal vision were recorded on mfERG2 and retested 1 year later (±15 days). The results showed that the mean (of all 103 local response measurements) implicit time *Z*-scores differed from 0.04 to 0.76 *Z*-scores with a mean difference of 0.36 ± 0.28 *Z*-scores. The amplitude *Z*-score differences ranged from 0.02 to 2.60 *Z*-scores, with a mean difference of 0.85 ± 0.81 *Z*-scores. For these 9 subjects, coefficients of variation (CV) were also calculated for the raw data of each of the 103 hexagons for both implicit time and amplitude. The local implicit time CVs ranged from 2.2% to 4.3% with a whole eye average of 3.0 ± 0.5%. The local amplitude CVs ranged from 10.5% to 47.3% with a whole eye average of 23.7 ± 6.9% (data not shown). This indicates that implicit time remains fairly stable over recording sessions but amplitudes are more variable.

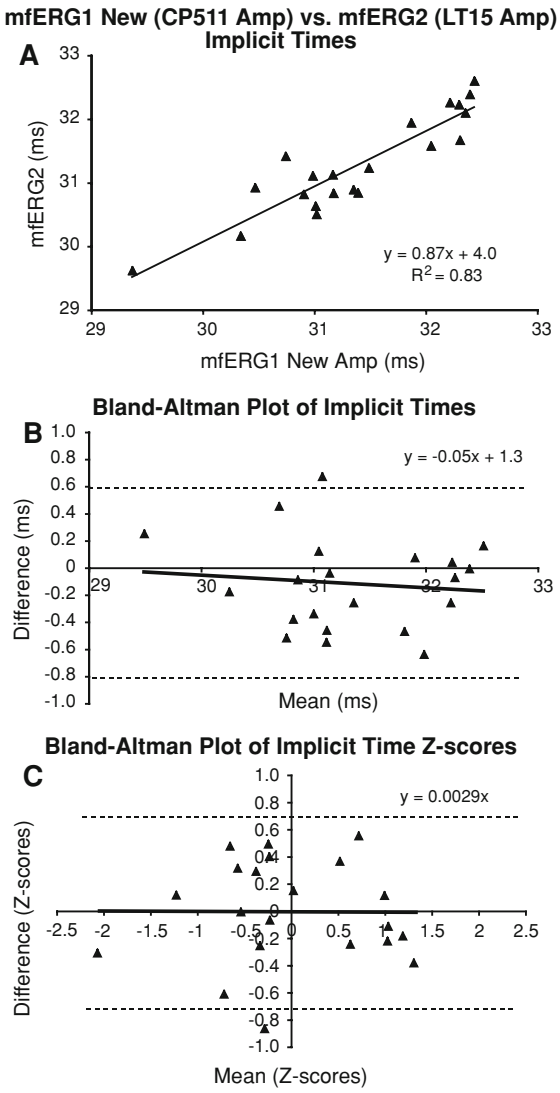

Fig. 5 Implicit time comparison for the newer amplifier of mfERG1 and mfERG2. **a** Comparison of the implicit times (ms). Each point is a whole eye average of one subject from group 3. **b** A Bland–Altman plot of that same data as **a**. The dashed lines on the plot indicate the 95% confidence interval and the solid line is the mean difference (0.1 ms) between the instruments for all 21 subjects. The slope of this line is not statistically different from zero ($P = 0.65$). **c** The Bland–Altman plot of the *Z*-scores of the implicit time data with the dashed lines indicating the 95% confidence interval and the solid line indicating the mean difference (0 *Z*-score units) for the 21 subjects. The slope of this line is not different from zero ($P = 0.98$)

The last analysis explored the similarity of implicit time and amplitude measures of the 103 *local* mfERGs obtained on the two instruments. The 95% confidence intervals of the difference between mfERG2 and the older amplifier of mfERG1 (the most different hardware configurations) were evaluated at all 103 retinal locations for the 21 subjects in group 3. The plots in Fig. 6 show the individual *Z*-score confidence intervals, represented as vertical gray bars, and the mean confidence intervals, represented as dashed horizontal lines. For implicit time, the mean local difference between the two instruments was 0.01 *Z*-score. The dashed horizontal lines in Fig. 6a indicate the mean 95% confidence interval (1.07 to −1.05). The locations near the blind spot (e.g., elements 48 and 59), are the most variable (up to 1.5 *Z*-scores in each direction). Based on these results, a local difference in implicit time must be greater than approximately 1 *Z*-score unit to differentiate it from inter-instrument variability and establish a significant functional change at a single retinal location. In this study, all of the subjects have normal vision (controls) and so no actual retinal defects existed. Both instruments agreed that all of these subjects were normal, with no subject having more than 4 local implicit time *Z*-scores ≥ 2.0 ($P = 0.91$). Figure 6b shows the 95% confidence intervals for the difference in amplitude *Z*-scores for the same responses. While the mean difference for all of the hexagons is small (0.04 *Z*-scores), the local amplitudes had more variation than the implicit times. The average 95% confidence interval for the amplitude *Z*-scores was 1.63 to −1.54 with some hexagons having a 95% confidence interval >2.0 *Z*-scores.

Discussion

The purpose of this study was to evaluate the robustness and stability of the mfERG as it is recorded over both time and with different instrumentation from the same manufacturer. In this study, we used VERIS^{TM} software and hardware. Although there have been several studies examining the repeatability and variability of the mfERG, the reproducibility of the mfERG across systems within a laboratory or across laboratories had not been examined. Understanding this reproducibility is a key component in pooling and comparing data across laboratories and replacing all or parts of an mfERG instrument during a study. For multicenter mfERG

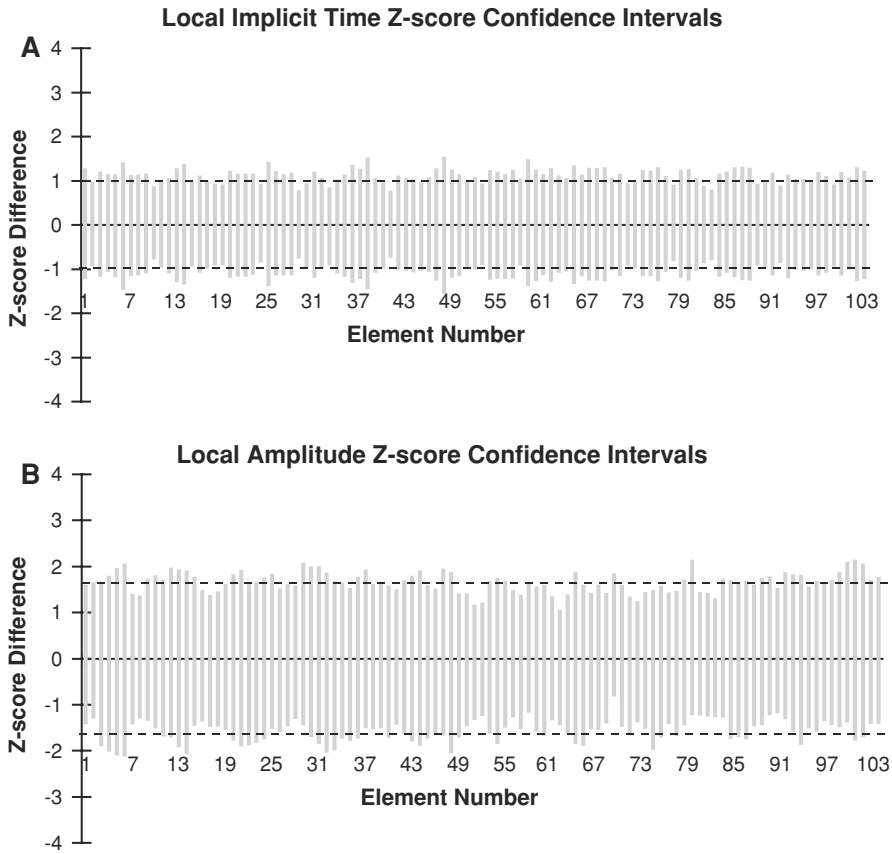

Fig. 6 **a** The 95% confidence intervals of the implicit time *Z*-score differences between the older amplifier of mfERG1 and mfERG2 for each of the 103 elements (gray vertical lines). The dashed lines indicate 1 *Z*-score in either direction, which is the average 95% confidence interval for all 103 hexagons. **b** The 95% confidence intervals of the amplitude *Z*-score difference between the older amplifier of mfERG1 and mfERG2 at each of the 103 hexagons. The dashed lines indicate 1.6 *Z*-scores, which is the average 95% confidence interval for this data

studies, the reproducibility, or agreement, of the response measures must be established first.

It is known from past studies that there are many sources of possible variation in the mfERG. It has been shown that differences in luminance [20], contrast [21], pupil size [22], adaptation states [23, 24], and even less than full correction of refractive error [25, 26] can all cause alterations in the mfERG. Furthermore, the way the data are filtered and processed during the recording session is another potential source of variability from session to session and laboratory to laboratory [27, 28]. These past studies have shown that while there are many factors that can cause variability, if they are controlled within a laboratory, the repeatability of the mfERG responses can be good, particularly with implicit time measures. All of these factors were controlled in this study in both intra-session and inter-session recordings. Furthermore, the use of the Hood and Li template scaling method in this study may have helped to improve reproducibility. Compared to measurements of peaks and troughs made manually, the template scaling method is more objective and less affected by noise. The method's relative insensitivity to noise is due to the fact that the waveform template is fit to the response being measured, using a least-squares criterion, over an 80 ms epoch. Thus, random noise in the region of the P1 peak has relatively little effect on either its estimated amplitude or implicit time.

Overall, we found the mfERG *Z*-scores for amplitude were satisfactorily reproducible and *Z*-scores for implicit time were very reproducible across time and with different instrumentation. The $\pm$ 0.86 *Z*-score confidence interval for mean implicit time corresponds to $\pm$ 0.73 ms, which is less than $\pm$ 1 real-time signal sample in our recordings. However, differences in recording instrumentation can cause raw response measures to be very different between instruments. These raw response differences can exist even when systems are similarly calibrated and when band-pass filter settings are nominally the same. In

our study, the amplifiers on mfERG1 were set to the same band-pass settings and records were taken simultaneously; the raw amplitudes were similar but raw implicit times were very different (2.5 ms mean difference) in the two channels. These implicit time differences are not surprising, given the different filter characteristics, but it must be noted that 2.5 ms is a large difference, more than 2.0 *Z*-scores. This difference is large enough to cause concern in a longitudinal study or comparison of data across laboratories, if one were not aware of the filtering differences between amplifiers. This difference could also lead to a belief that an eye had improved or deteriorated even when there was no actual change.

By normalizing, using data from a normal population, mfERG measurements are more comparable and in reasonably close agreement across instruments, and the effects of differences in instrumentation are minimized. The normal subject samples should be similar and matched appropriately to the disease state and patient sample being studied, as was the case in this study. There are multiple normalization methods, including percentiles and *Z*-scores. We chose *Z*-scores for a number of reasons. They include the mean and variability of the normative data and so they can be quickly used to identify abnormalities. However, most importantly, they transform the measurements so that they are relative to the control data collected on specific instruments. Another possible approach to making data more comparable is to band-pass filter recordings over a larger frequency range and then digitally filter the responses. This would likely remove some of the differences we observed in implicit time. Since digital filtering uses Fourier analysis, there is no phase shift as there can be in analog filtering. However, digital filtering would likely not help in making amplitude data more reproducible.

For the first part of our study, we performed amplifier comparisons using parallel channels. We did this to avoid time-varying (test–retest) factors and to isolate differences in the instrumentation. When comparing both the same and different instruments across time in the second part of our study, we found that amplitudes were much less repeatable than implicit times. This is in agreement with earlier studies, which have also found amplitudes to be more variable [29]. The CVs we found for both amplitudes and implicit times are in agreement with previous studies [12, 13, 15] when averaging over the whole eye. We also looked at CVs on a local level and found them to be fairly consistent across the retina when examining implicit time but highly variable for amplitudes. No CVs were calculated for *Z*-score data as CVs are poor estimates of variation when the mean of the data is near zero, which is the case for *Z*-scores of subjects with normal vision. However, the range of *Z*-score differences in amplitude measurements are also much more variable than it is for implicit times.

In general, comparison of different instruments involves true instrument differences (e.g., the hardware and software design) and test–retest variation. It appears that a large part of the variability between instruments that we observed, especially in amplitude, might come from inter-session rather then inter-instrument sources. Most of the response variation we observed between the instruments, after conversion to *Z*-scores, was of the same magnitude as test–retest on the same instrument with the same amplifier. Therefore, it appears that data collected on different setups can be compared more easily after conversion to *Z*-scores, at least when recording conditions are sufficiently equated.

Previous studies examining the repeatability of the mfERG have typically used ring averages to look at the differences between different sessions. This study uses comparisons among eye averages and also among local response measurements. In agreement with other studies [12, 13], we found, not surprisingly, that the local measures are less repeatable in comparison to whole eye averages. There are a number of reasons why local measurements can be less repeatable than eye averages, including a lower signal to noise ratio, small changes in stimulus placement on the retina, and changes in electrode placement in the case of amplitudes.

In conclusion, the mfERG is quite reproducible, even across different recording installations. This study suggests that it is possible to compare and/or combine data obtained from different instrumentation, provided that sufficiently large and similar normative data sets are collected on each instrument. Conversion of raw mfERG measurements to *Z*-scores based on normative data is an efficient and effective means to compare or combine measurements obtained with different instrumentation. Such comparisons and combinations are critical to multicenter studies, some longitudinal studies, and to following patients over years of care.

Acknowledgments This study was supported an American Optometric Foundation Ezell Fellowship (WWH) and NIH/NEI grants (T32 EY007043-28 and R01 EY02271 to AJA).

References

1. Parisi V, Perillo L, Tedeschi M, Scassa C, Gallinaro G, Capaldo N et al (2007) Macular function in eyes with early age-related macular degeneration with or without contralateral late age-related macular degeneration. Retina 27:879–890. doi:10.1097/01.iae.0000256039.59142.22
2. Janaky M, Palffy A, Deak A, Szilagyi M, Benedek G (2007) Multifocal ERG reveals several patterns of cone degeneration in retinitis pigmentosa with concentric narrowing of the visual field. Invest Ophthalmol Vis Sci 48:383–389. doi:10.1167/iovs.06-0661
3. Han Y, Bearse MA Jr, Schneck ME, Barez S, Jacobsen CH, Adams AJ (2004) Multifocal electroretinogram delays predict sites of subsequent diabetic retinopathy. Invest Ophthalmol Vis Sci 45:948–954. doi:10.1167/iovs.03-1101
4. Han Y, Schneck ME, Bearse MA Jr, Barez S, Jacobsen CH, Jewell NP et al (2004) Formulation and evaluation of a predictive model to identify the sites of future diabetic retinopathy. Invest Ophthalmol Vis Sci 45:4106–4112. doi:10.1167/iovs.04-0405
5. Lai TY, Chan WM, Lai RY, Ngai JW, Li H, Lam DS (2007) The clinical applications of multifocal electroretinography: a systematic review. Surv Ophthalmol 52:61–96. doi:10.1016/j.survophthal.2006.10.005
6. Ng JS, Bearse MA Jr, Schneck ME, Barez S, Adams AJ (2008) Local diabetic retinopathy prediction by multifocal ERG delays over 3 years. Invest Ophthalmol Vis Sci 49:1622–1628. doi:10.1167/iovs.07-1157
7. Lyons JS, Severns ML (2007) Detection of early hydroxychloroquine retinal toxicity enhanced by ring ratio analysis of multifocal electroretinography. Am J Ophthalmol 143:801–809. doi:10.1016/j.ajo.2006.12.042
8. Kardon RH, Morrisey MC, Lee AG (2006) Abnormal multifocal electroretinogram (mfERG) in ethambutol toxicity. Semin Ophthalmol 21:215–222. doi:10.1080/08820530600987454
9. Tari SR, Vidne-Hay O, Greenstein VC, Barile GR, Hood DC, Chang S (2007) Functional and structural measurements for the assessment of internal limiting membrane peeling in idiopathic macular pucker. Retina 27:567–572. doi:10.1097/IAE.0b013e31802ea53d
10. Schatz P, Holm K, Andreasson S (2007) Retinal function after scleral buckling for recent onset rhegmatogenous retinal detachment: assessment with electroretinography and optical coherence tomography. Retina 27:30–36. doi:10.1097/01.iae.0000256659.71864.83
11. Meigen T, Friedrich A (2002) The reproducibility of multifocal ERG recordings. Ophthalmologe 99:713–718. doi:10.1007/s00347-002-0630-0
12. Parks S, Keating D, Williamson TH, Evans AL, Elliott AT, Jay JL (1996) Functional imaging of the retina using the multifocal electroretinogram: a control study. Br J Ophthalmol 80:831–834. doi:10.1136/bjo.80.9.831
13. Gundogan FC, Sobaci G, Bayraktar MZ (2008) Intra-sessional and inter-sessional variability of multifocal electroretinogram. Doc Ophthalmol
14. Yoshii M, Yanashima K, Wakaguri T, Sakemi F, Kikuchi Y, Suzuki S et al (2000) A basic investigation of multifocal electroretinogram: reproducibility and effect of luminance. Jpn J Ophthalmol 44:122–127. doi:10.1016/S0021-5155(99)00189-6
15. Bultmann S, Rohrschneider K (2002) Reproducibility of multifocal ERG using the scanning laser ophthalmoscope. Graefes Arch Clin Exp Ophthalmol 240:841–845. doi:10.1007/s00417-002-0564-x
16. Hood DC, Bach M, Brigell M, Keating D, Kondo M, Lyons JS et al (2008) ISCEV guidelines for clinical multifocal electroretinography (2007 edition). Doc Ophthalmol 116:1–11. doi:10.1007/s10633-007-9089-2
17. Hood DC, Li J (1997) A technique for measuring individual multifocal ERG records. Trends Opt Photon 11:280–293
18. Bearse MA Jr, Adams AJ, Han Y, Schneck ME, Ng J, Bronson-Castain K et al (2006) A multifocal electroretinogram model predicting the development of diabetic retinopathy. Prog Retin Eye Res 25:425–448. doi:10.1016/j.preteyeres.2006.07.001
19. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1:307–310
20. Schimitzek T, Bach M (2006) The influence of luminance on the multifocal ERG. Doc Ophthalmol 113:187–192. doi:10.1007/s10633-006-9028-7
21. Tam A, Chan H, Brown B, Yap M (2004) The effects of forward light scattering on the multifocal electroretinogram. Curr Eye Res 28:63–72. doi:10.1076/ceyr.28.1.63.23494
22. Gonzalez P, Parks S, Dolan F, Keating D (2004) The effects of pupil size on the multifocal electroretinogram. Doc Ophthalmol 109:67–72. doi:10.1007/s10633-004-1545-7
23. Chappelow AV, Marmor MF (2002) Effects of pre-adaptation conditions and ambient room lighting on the multifocal ERG. Doc Ophthalmol 105:23–31. doi:10.1023/A:1015713029443
24. Chen JC, Brown B, Schmid KL (2006) Changes in implicit time of the multifocal electroretinogram response following contrast adaptation. Curr Eye Res 31:549–556. doi:10.1080/02713680600744869
25. Chan HL, Siu AW (2003) Effect of optical defocus on multifocal ERG responses. Clin Exp Optom 86:317–322
26. Pieh C, Hoffmann MB, Bach M (2005) The influence of defocus on multifocal visual evoked potentials. Graefes Arch Clin Exp Ophthalmol 243:38–42. doi:10.1007/s00417-004-0969-9
27. Han Y, Bearse MA Jr, Schneck ME, Barez S, Jacobsen C, Adams AJ (2004) Towards optimal filtering of "standard"

multifocal electroretinogram (mfERG) recordings: findings in normal and diabetic subjects. Br J Ophthalmol 88:543–550. doi:10.1136/bjo.2003.026625

28. Oyamada MK, Dotto Pde F, Abdalla M (2007) Technical factors that influence multifocal electroretinogram (mfERG) recording. Arq Bras Oftalmol 70:713–717. doi:10.1590/S0004-27492007000400027

29. Fortune B, Schneck ME, Adams AJ (1999) Multifocal electroretinogram delays reveal local retinal dysfunction in early diabetic retinopathy. Invest Ophthalmol Vis Sci 40:2638–2651