Test–Retest Variability of Multifocal Visual Evoked Potential and SITA Standard Perimetry in Glaucoma

Anne Bjerre, John R. Grigg, Neil R. A. Parry, and David B. Henson

PURPOSE. To investigate the test-retest variability of multifocal visual evoked potential (mfVEP) and threshold perimetry in glaucoma, and to examine the relationship between the two techniques.

METHODS. Data were recorded using the AccuMap mfVEP and SITA standard program of the Humphrey Field Analyzer. Data were obtained twice within a 4-week period from both eyes of 74 patients with varying amounts of glaucomatous visual field loss. The number of defective test locations (those falling beyond a given probability value of being normal) were calculated for mfVEP and SITA, using databases incorporated within the instruments software. Reliability measures and test times were recorded along with patient test preference.

RESULTS. Both tests showed a large degree of test-retest variability in the number of defective test locations (95% limits of agreement for mfVEP and SITA being 13.39 and 9.88, respectively). A “fair to moderate” degree of spatial agreement was found between mfVEP and SITA. The number of mfVEP defective locations was dependent on the signal amplitude. No relationship was found between test-retest variability and the reliability indices for either test. The mean time taken to perform mfVEP and SITA standard was 53 and 20 minutes, respectively, and 73 of the 74 patients preferred the mfVEP test.

CONCLUSIONS. Test-retest variability was found to be slightly greater for mfVEP. The processing of mfVEP signals needs to be changed to remove the relationship between the number of defective locations and signal amplitude. The majority of patients preferred mfVEP to conventional perimetry although mfVEP takes longer to perform. (Invest Ophthalmol Vis Sci. 2004;45:4035–4040) DOI:10.1167/iovs.04-0099

Although threshold perimetry is widely used to diagnose and monitor glaucoma, the test has major shortcomings. The results from threshold perimetry are variable, making it difficult to differentiate between real change and random fluctuation, which can lead to delays in the detection of change in the visual field. Threshold perimetry is also time consuming, is fatiguing for the patient, and shows a significant learning effect.1-3 These shortcomings have led clinicians to seek alternative ways of detecting and monitoring glaucoma. One of these is the multifocal visual evoked potential (mfVEP).

The mfVEP technique simultaneously records from many spatially localized, cortically scaled stimuli. Early problems of large intersubject variability have been reduced with improved electrode placements,4 electroencephalogram (EEG) scaling,5 and asymmetry analysis between eyes.6,7 These improvements to the mfVEP technique have been incorporated in the multifocal objective perimeter (AccuMap; ObjectiVision Pty Ltd, Sydney, Australia).

The early detection of glaucomatous progression relies on both the sensitivity and reproducibility of measures. Proposed alternatives to threshold perimetry need to demonstrate improvements in these areas.

The Swedish Interactive Threshold Algorithms (SITA) have recently been developed for the Humphrey Field Analyzer (HFA model 740; Carl Zeiss Meditec, Dublin, CA) by Bengtsson et al.8 These strategies offer several advantages over earlier threshold strategies, notably a reduction in test time9-14 and a small reduction in test-retest variability.15

This study assessed the role of the mfVEP in monitoring glaucoma by comparing the test-retest variability of mfVEP to SITA standard perimetry. Variability was well defined by test-retest trials where a population of glaucomatous patients was examined on 2 separate days with the same test. The study further aimed to look at the relationship between results obtained with mfVEP and SITA standard, to report on test times, and patient preference.

METHODS

Patients

Patients attending Manchester Royal Eye Hospital with a diagnosis of glaucoma were recruited for the study. Inclusion criteria were visual field loss in at least one eye (Glaucma Hemifield Test outside normal limits), age greater than 40 years, spherical refractive error within ± 5.00 D and the cylinder component within ± 3.00 DC, and corrected visual acuity of 0.5 LogMAR or better. Patients with narrow-angle glaucoma, secondary glaucoma, and ocular pathologies other than early cataract were excluded. The sample was designed to be as representative as possible of glaucoma patients attending the hospital.

Patients were not excluded on the basis of perimetric reliability criteria, although all patients had prior experience of static automated threshold perimetry (none of the patients had prior experience of mfVEP before the study).

The study adhered to the tenets of the Declaration of Helsinki for research involving human subjects and was approved by the Central Manchester Research Ethics Committee. Informed consent was obtained from all the participants before enrollment in the study.

Multifocal Visual Evoked Potential (mfVEP) Technique

The mfVEP data was recorded using the AccuMap instrument (Version 1.0; ObjectiVision Pty Ltd, Sydney, Australia). This instrument uses a spread spectrum technique with families of binary sequences to drive the visual stimulus. The stimulus is presented as a pseudorandom, cortically scaled pattern on a computer screen (22-inch high-resolution display; Hitachi, Tokyo, Japan). There were 58 test locations, with each location containing a checkerboard pattern of 16 checks. Eight of the...
checks were white with a luminance of 146 cd/m², and the other eight were black with a luminance of 1.1 cd/m², generating a Michelson contrast of 99%. The background luminance of the screen was 73.5 cd/m². Fifty-six of the test locations were within 24° of eccentricity and the remaining two were located in the nasal region within 32° of eccentricity. The stimulus was driven at a frame rate of 75 Hz. The central area of 1° was used as a fixation target. A series of randomly changing numbers was presented at the fixation point and the patient was requested to press a button when the number 3 was seen. This was designed to enhance attention and fixation, and was used to provide indices for test reliability.

During each run, the VEP amplitudes for all the test locations were recorded for 55 seconds. EEG scaling of the VEP amplitude was automatically carried out to reduce intersubject variability. The patient was seated in a dimly lit room at a distance of 30 cm from the computer screen, with the chin slightly elevated to relax the neck muscles. A full refractive correction for near vision was worn and pupils were not dilated. Four gold disc electrodes (Grass; Astro-Med, Inc., West Warwick, RI) placed in a custom designed occipital cross electrode holder were used to permit four-channel bipolar recording. The vertical channel received information from two electrodes positioned 2.5 cm above and 4.5 cm below the inion. The horizontal channel obtained information from two electrodes located 4 cm either side of the inion. The two oblique channels received input from the lower midline electrode and either right or left horizontal electrode. The scalp was cleaned at the site of each electrode using Nuprep (D.O. Weaver and Company, Aurora, CO), and contact gel (Skintact ECG Gel; Leonhard Lang, Austria) was applied between the scalp and electrodes. The impedance of each electrode was measured with a target value of 5 kΩ or less. Raw trace data for each channel was presented in real time during each run. EEG scaling of the VEP amplitude using fast Fourier analysis is automatically carried out after each run to reduce intersubject variability. A Fourier spectrum window display is used to detect any high alpha component or electrocardiogram (ECG) contribution after each run. The analysis software incorporates a trace improvement algorithm, which is an index of the signal-to-noise ratio. The system was repeatedly run, with noisy runs replaced, until a minimum of seven good quality runs was collected from each eye.

Visual field data was collected with a Humphrey Field Analyzer (Carl Zeiss Meditec) using program 24–2, SITA standard. The test was performed with the appropriate refractive correction in large aperture lenses specifically designed for perimetry. All mfVEP and visual field data were collected by the same experienced examiner. Each patient was seen twice within a 38-day period. The sequence of the two tests was randomized between patients but constant for each patient.

Analysis

The analysis was based on the number (maximum of 52 for the HFA, excluding 2 points falling within the region of the blind spot, and 58 for the mfVEP) and spatial location of the test points marked as outside of normal limits (5%, 2%, and 1%) by each instrument’s software. The mfVEP reliability indices (fixation losses [FL] and false positives [FP]) and the visual field reliability indices (FL, FP, and false negatives [FN]), were analyzed.

The time taken to perform the mfVEP was recorded in a subset of patients (n = 24). To determine which test the patients preferred, all patients were asked two standardized questions by the examiner when they had completed both tests on the first visit: “Which test did you prefer?” and “Why did you prefer this test?”

Results

Eighty-two patients were enrolled in the study. Six patients failed to attend both sessions and two patients were excluded.
because of poor fixation or headshaking, leaving a residual 
number of test locations falling beyond the 5% level of normal-
ity for mfVEP and SITA tests are given in Figures 1 and 2. These 
figures show a significant amount of test-retest variability for 
both tests with the limits of agreement being slightly larger 
for the mfVEP test than SITA perimetry (13.39 and 9.88, respec-
tively). This difference reversed when higher cut-off levels (2% 
and 1%) were used (Table 1). Prior experience with threshold 
perimetry, but not mfVEP, might have influenced the measured 
variability with the two tests. 
The inclusion criteria of visual field loss (Glaucoma Hemi-
field Test outside normal limits) on the first visit, accounts for 
the absence of patients with zero defective locations on visit 1, 
but not on visit 2. This may have introduced a small bias in the 
data.

**Test–Retest Variability**

Test-retest variability and the 95% limits of agreement for 
the number of test locations falling beyond the 5% level of normal-
ity for mfVEP and SITA tests are given in Figures 1 and 2. These 
figures show a significant amount of test-retest variability for 
both tests with the limits of agreement being slightly larger 
for the mfVEP test than SITA perimetry (13.39 and 9.88, respec-
tively). This difference reversed when higher cut-off levels (2% 
and 1%) were used (Table 1). Prior experience with threshold 
perimetry, but not mfVEP, might have influenced the measured 
variability with the two tests. 
The inclusion criteria of visual field loss (Glaucoma Hemi-
field Test outside normal limits) on the first visit, accounts for 
the absence of patients with zero defective locations on visit 1, 
but not on visit 2. This may have introduced a small bias in the 
data.

**Quadrant and Hemifield Spatial Agreement**

The spatial agreement between SITA and mfVEP data was 
investigated with two by two contingency tables (Table 2), for 
both the four quadrants and the two hemifields. The cut-off 
criteria for the defective classification was ≥3 and ≥5 test 
locations beyond the 5% level of normality for quadrants and 
hemifield, respectively. These values were chosen to give spec-
ficities of less than, but close to 2%. For comparison purposes, 
similar contingency tables were compiled for visits 1 and 2 of 
the SITA and mfVEP data. The percentage levels of agreement 
and kappa values are given in Table 3.

Whereas the levels of agreement between mfVEP and SITA 
appear to be fairly good at 70% and 75%, the chance corrected 

**Table 1. The Limits of Agreement at the 95, 97 and 99 Probability Levels for mfVEP and SITA Standard**

<table>
<thead>
<tr>
<th>P Value</th>
<th>mfVEP</th>
<th>SITA Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>95%</td>
<td>13.39</td>
<td>9.88</td>
</tr>
<tr>
<td>97%</td>
<td>9.97</td>
<td>8.99</td>
</tr>
<tr>
<td>99%</td>
<td>6.78</td>
<td>9.28</td>
</tr>
</tbody>
</table>

**Table 2. Two by Two Contingency Table Showing Level of Spatial Agreement between mfVEP and SITA**

<table>
<thead>
<tr>
<th>SITA Standard</th>
<th>Normal</th>
<th>Defective</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfVEP</td>
<td>133</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>108</td>
<td>281</td>
</tr>
<tr>
<td></td>
<td>241</td>
<td>351</td>
</tr>
</tbody>
</table>

Data from the four quadrants has been combined with ≥3 test 
locations lying beyond the 5% level of normality being classified as 
defective. Agreement = (133 + 281)/592 = 70%.

**Table 3. Levels of Agreement and Kappa Values for mfVEP and SITA Perimetry**

<table>
<thead>
<tr>
<th>Quadrants</th>
<th>Agreement</th>
<th>Kappa</th>
<th>Hemifields</th>
<th>Agreement</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>mfVEP visit 1 versus SITA visit 1</td>
<td>70%</td>
<td>0.33</td>
<td>75%</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>SITA visit 1 versus SITA visit 2</td>
<td>83%</td>
<td>0.65</td>
<td>85%</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>mfVEP visit 1 versus mfVEP visit 2</td>
<td>78%</td>
<td>0.66</td>
<td>86%</td>
<td>0.63</td>
<td></td>
</tr>
</tbody>
</table>

Data from all four quadrants and both hemifields has been com-
bined.

kappa values of agreement indicate that there is only “fair” 
agreement (0.21–0.40) between the two tests. In compari-
on, the test-retest levels of agreement were higher with 
kappa values falling in the “good” agreement category (0.61-
0.80).19

**Relationship between Extents of Loss**

A poor relationship was found between the number of test 
locations marked as beyond the 5% level of normality with the 
mfVEP and SITA tests (Fig. 3). This relationship was much 
bad than that for the test-retest data of either test (see Figs. 
1 and 2). Numerous cases were found to have a large number 
of defective locations with one test but mild or no loss with the 
other test (see Fig. 4 for examples of good and poor corre-
spondence between the mfVEP and SITA tests). 
The mfVEP test, on average, classified more locations as 
being beyond the 95% level of normality than SITA (mean 
difference 4.64 ± 14.62, Fig. 5), although the difference was 
small and reduced further at higher cut off levels (98% level, 
mean 2.93 ± 12.03). 

These differences between the two tests were often large 
and difficult to explain on the basis of test-retest variability. 
To investigate this point further, the differences were compared 
between visits 1 and 2. Figure 6 shows that the differences 
between the two tests are consistent and are not, therefore, 
due to test-retest variability.

To investigate whether this finding was related to the am-
plitude of the mfVEP signals, the differences between SITA 
perimetry and mfVEP were compared to the summed ampli-
tude of the mfVEP data (Fig. 7). This figure shows a strong 
relationship (P < 0.0001) between these two measures, as the 
summed amplitude of the mfVEP signal reduces the mfVEP 
results show more defective test locations than the visual field 
test.

**Reliability Indices and Test Times**

The reliability indices for SITA perimetry showed no statisti-
cally significant relationship to test-retest variability, (FL: P = 
0.289; FP: P = 0.052; EN: P = 0.300). A poor relationship was 
also found between the mfVEP test and its reliability indices 
(FL: P = 0.349; FP: P = 0.698).

The mean time (± SD) taken to perform mfVEP on both 
eyes was 33.60 ± 3.02 minutes while SITA perimetry took 
approximately 20 minutes (~8 minutes per eye in cases of 
visual field loss15 with 4 minutes setup/change over time). 
Despite the time difference and the fact that the mfVEP test is 
more invasive, 73 of the 74 patients preferred the mfVEP test. 
The main stated reasons were that the test was easier and less 
stressful. Many individuals found conventional perimetry bor-
ing and confusing. Physical comfort was also a factor for many 
patients.
DISCUSSION

Our test–retest measures of SITA perimetry show a large degree of variability, a finding which is in agreement with previous studies.15,21–23 This large degree of variability in patients with prior experience of threshold perimetry limits the capacity of current perimetric techniques to quantify change and provide accurate predictions of long-term outcomes. This shortcoming of current perimetric techniques has encouraged the development of alternative technologies such as the mfVEP. This study found that the test–retest variability of mfVEP was slightly larger than SITA perimetry in a sample of patients with glaucomatous visual field loss and no prior experience of mfVEP. The current mfVEP technique does not, therefore, seem to offer any advantages over current perimetric techniques when it come to quantifying change and predicting long-term outcomes.

Goldberg et al.16 examined the coefficient of variation for the amplitude of the mfVEP in normal subjects, and the same research group investigated the coefficient of variation in glaucoma suspects and confirmed cases of glaucoma (Graham SL, et al. IOVS 2003;44:ARVO E-Abstract 45). The glaucoma patients demonstrated significantly more variability than the normal subjects and suspect glaucoma patients. The variability was higher in areas of smaller signal amplitude and the authors concluded that a large change in the amplitude is needed to confirm progression of a visual field defect. Furthermore, they

![Figure 3](image1)

**Figure 3.** Relationship and 95% prediction limits for the number of test locations falling beyond the 95% level of normality between SITA and mfVEP. Data taken from visit 1.

![Figure 4](image2)

**Figure 4.** (A) Example of good correspondence between the mfVEP and SITA results. Reliability indices for both tests were <10%. (B) Example of poor correspondence between the mfVEP and SITA perimetry. The mfVEP pattern deviation probability plot shows a larger number of abnormal test locations compared to SITA. Reliability indices for both tests were <10%.
suggested improved signal-to-noise ratio should improve reproducibility.

The spatial agreement between mfVEP and SITA perimetry was similar to that recently reported by Hood et al.24 Agreement was generally in the region of 70%–75%, but when corrected for chance using the kappa statistic was found to be only “fair to moderate.”19 For comparison the agreement between repeat measures with the same test (mfVEP and SITA perimetry) was found by the kappa statistic to be “good”.

An unexpected finding of this study was the occasional large disparity between the number of test locations falling beyond the 95% level of normality with the two tests. This disparity was not consistently in one direction. In some patients the mfVEP marked more locations as being abnormal than SITA perimetry, whereas with other patients it was the other way around. Similar findings have been reported by other researchers.7,24–26 Further investigation of these disparities between mfVEP and SITA perimetry revealed that they were consistent from one session to another (i.e., if a patient had more defective locations with mfVEP at visit 1, then a similar result would be obtained at visit 2). This finding indicated that the differences between the two tests could not be explained on the basis of random variability. Further analysis of the differences between the two tests revealed a relationship with the amplitude (peak-to-trough) of the VEP signal. In cases where the amplitude was low, the mfVEP tended to classify more test locations as being abnormal than did SITA perimetry. When the VEP amplitude was high, the mfVEP test classified fewer locations as being abnormal.

The signal-to-noise ratio of mfVEP varies from subject to subject and has been highlighted by other research groups as an important parameter of test performance.24,27,28 Changes in electrode resistance and cortical architecture are two of the parameters that influence this ratio. To compensate for differences in the signal-to-noise ratio, the mfVEP signal undergoes a certain amount of processing. This processing should be independent of defect size. The results of the present study indicate that this is not the case and that a negative relationship exists between signal amplitude and the number of locations classified as outside normal limits.

Visual field reliability estimates have previously been shown to have low reliability.29–31 The poor relationship between visual field reliability indices and test–retest variability found in this study is not, therefore, surprising. It does, however, highlight the inadequacy of reliability indices at providing the clinician with useful information regarding test–retest variabil-

![Figure 5](https://example.com/f5.png)

**FIGURE 5.** Distribution of the differences between the number of test locations that fell beyond the 95% limits of normality for the mfVEP and SITA tests.

![Figure 6](https://example.com/f6.png)

**FIGURE 6.** The relationship and 95% prediction limits between visit 1 and visit 2 of the differences in the number of test locations that fell beyond the 95% limits of normality for the mfVEP and SITA tests.

![Figure 7](https://example.com/f7.png)

**FIGURE 7.** The relationship and 95% prediction limits between the summed amplitude of the mfVEP signal (peak-to-trough, within 60 to 180 ms window) and the difference between the number of test locations classified as beyond the 95% level of normality with the mfVEP and SITA perimetry.
ity. The mfVEP reliability estimates FL and FP are derived from the patients’ responses to the fixation target. Patients are instructed at the onset of the test to press a response button every time a number 3 appears at the fixation point. If the patient fails to press the response button when a number 3 is presented, this is recorded as a FL. If the patient presses the button when a different number is displayed, this is recorded as a FP. While both FPs and FLs were found to increase with the number of defective test locations, there was little relationship between these measures of reliability and test–retest performance. This again signifies that these measures of reliability have little clinical value. The use of an infrared video camera that monitors the patient’s eye movements could give valuable information about fixation accuracy. None of the patients were unable to perform the visual field test, although this might simply reflect the fact that one of the inclusion criteria was prior experience with the HFA. Two patients had to be excluded from the study because the mfVEP responses were very noisy. One patient was constantly head-shaking and the other patient was having difficulty maintaining fixation due to a visual acuity of +0.5 LogMAR. Recently, it was highlighted that cataract, uncorrected refractive error, and unstable fixation can produce apparent mfVEP defects (Winn BJ, et al. IOVS 2003;44:ARVO E-Abstract 52).

Multifocal VEP is a relatively new technology compared with perimetry and as such is still evolving at a fairly rapid rate. It should be possible to modify the mfVEP software to remove the relationship between the number of locations classed as defective and the signal amplitude or, at least, to highlight when the signal amplitude is likely to lead to erroneous results. Test–retest variability needs to be improved if mfVEP is going to provide a more reliable measure of change and a better predictor of long term outcomes than SITA perimetry.

References


