Comparing Multifocal VEP and Standard Automated Perimetry in High-Risk Ocular Hypertension and Early Glaucoma

Brad Fortune,[1] *Shaban Demirel,*[1] *Xian Zhang,*[2] *Donald C. Hood,*[2] *Emily Patterson,*[1] *Annisa Jamil,*[1] *Steven L. Mansberger,*[1] *George A. Cioffi,*[1] *and Chris A. Johnson*[1]

Purpose. To compare the diagnostic performance of multifocal visual evoked potential (mfVEP) and standard automated perimetry (SAP), in eyes with high-risk ocular hypertension or early glaucoma.

Methods. Both eyes of 185 individuals with high-risk ocular hypertension or early glaucoma were evaluated. Subjects ranged in age from 37 to 87 (mean ± SD: 61 ± 11 years). Pattern-reversal mfVEPs were obtained by using VERIS (Electro-Diagnostic Imaging, San Mateo, CA) with a four-electrode array and were analyzed with custom software. SAP visual fields (SITA-standard; Carl Zeiss Meditec, Inc., Dublin, CA) were obtained within 22.3 ± 27.0 days of the mfVEP. Stereo disc photographs and Heidelberg Retina Tomograph (HRT) images were obtained during one visit, which was within 24.8 ± 50.4 days of the mfVEP and 33.1 ± 62.9 days of the SAP visual field. Abnormalities on the mfVEP were defined by using a variety of cluster criteria: SAP with pattern standard deviation (PSD) $P \leq 0.05$ or glaucoma hemifield test (GHT) outside normal limits, according to OHTS criteria (SAP-OHTS). In separate analyses cluster criteria were used to determine SAP abnormalities. Disc photographs were graded as either glaucomatous optic neuropathy (GON) or normal by two independent masked experts, and disagreements were adjudicated by a third masked expert. The overall Moorfields regression analysis (MRA) result from the HRT was used as a separate diagnostic classification. All eyes classified as "borderline" by the MRA were assigned to the normal category (i.e., "within normal limits"). Sensitivity for mfVEP or SAP was defined as the percentage of GON eyes that had an abnormality on the functional test. Specificity for mfVEP or SAP was defined as the percentage of eyes with normal optic disc structure that had normal functional test results.

Results. Disc photographs from 50% of eyes were graded GON. Both eyes were graded GON in 71 (38%) of the 185 subjects. Exactly half as many eyes were abnormal by HRT MRA. The average SAP mean deviation (MD) was +0.3 ± 2.1 dB; average PSD was 2.3 ± 1.9 dB. By OHTS criteria, 83 (22%) of the 370 eyes had an abnormal SAP. Depending on the cluster criterion used, the proportion of eyes with an abnormal SAP ranged from 8% to 26% and with an abnormal mfVEP, from 14% to 45%. A criterion with an estimated specificity in normal subjects of 91% resulted in 102 (28%) eyes with an abnormal mfVEP. For criteria with estimated specificities of 95% and 99%, respectively, 88 (24%) eyes and 52 (14%) eyes had an abnormal mfVEP. Agreement between SAP and mfVEP ranged from 75% to 81%. The sensitivity of SAP-OHTS to detect GON (using the disc photograph as diagnostic standard) was 29%, whereas specificity was 84%. Sensitivity of the mfVEP to detect GON, for cluster criteria with disc structure specificity between 84% and 87%, ranged from 28% to 32%. When the HRT MRA was used as the diagnostic standard, sensitivities of both functional tests to detect GON increased to 42%.

Conclusions. The diagnostic performance of mfVEP was similar to that of SAP. However, the two modalities agreed in only ~80% of eyes, suggesting that they may detect slightly different functional deficits. (*Invest Ophthalmol Vis Sci.* 2007;48: 1173–1180) DOI:10.1167/iovs.06-0561

The multifocal visual evoked potential (mfVEP) has steadily gained acceptance as an objective method for topographical assessment of visual function in glaucoma.[1–14] Improvements on the fundamental technique,[15] such as multiple channel recordings,[2,16] interocular analyses,[3,4] and signal-to-noise amplitude assessment,[17,18] have helped to increase the diagnostic utility of the mfVEP. After these improvements, several studies have demonstrated that the mfVEP can accurately discriminate between persons with glaucomatous visual field loss and those with normal vision.[8–10,13,14] In fact, some studies have shown that the mfVEP can also detect abnormalities in unaffected areas of glaucomatous visual fields,[10,13] or even in persons with suspected glaucoma whose visual fields are completely within normal limits.[2,3,8,14] The latter findings pose a challenge because they may represent greater sensitivity of the mfVEP to detect early functional loss, or alternatively, a higher false-alarm rate. Longitudinal studies will help to determine which of these interpretations is most accurate.

As mentioned, several of these studies have directly compared the diagnostic performance of the mfVEP with the most widely used method of functional assessment, namely standard automated perimetry (SAP). Yet all but one of those comparisons were confounded by the use of SAP as the diagnostic gold standard, at least in part, to define the study populations. This problem was perhaps best elucidated by Bowd et al.[19] in their discussion regarding the effects of using an imperfect standard to evaluate another diagnostic test. Graham et al.[14] recognized this problem and designed a portion of their retrospective study on the mfVEP so that the structural appearance of the optic disc (as judged by masked experts), could be used as the diagnostic standard. They found the performance of mfVEP and SAP to be quite similar.[14] A theoretical analysis by Hood et al.[11,13,20] predicted that the performance of a monocular mfVEP test would be approximately equal to that of SAP and

From [1]Discoveries in Sight, Devers Eye Institute, Portland, Oregon; and the [2]Department of Psychology, Columbia University, New York, New York.

Supported by the M. J. Murdock Charitable Trust, Vancouver WA; Good Samaritan Foundation, Portland, OR; National Eye Institute Grants R01-EY03424 (CAJ) and R01-EY02115 (DCH); and the Legacy Good Samaritan Foundation.

Submitted for publication May 22, 2006; revised October 4 and 26, 2006; accepted January 15, 2007.

Disclosure: **B. Fortune**, None; **S. Demirel**, None; **X. Zhang**, None; **D.C. Hood**, None; **E. Patterson**, None; **A. Jamil**, None; **S.L. Mansberger**, None; **G.A. Cioffi**, None; **C.A. Johnson**, None

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Corresponding author: Brad Fortune, Discoveries in Sight, Devers Eye Institute, Legacy Clinical Research and Technology Center, 1225 NE Second Avenue, Portland, OR 97232; bfortune@deverseye.org.

would be slightly better than SAP when the interocular analysis was added to the mfVEP.

The primary purpose of the present study was to compare the diagnostic performance of mfVEP and SAP, in eyes with high-risk ocular hypertension or early glaucoma. Two different characterizations of optic disc structure were used as the diagnostic standard on which the comparison of these functional measures was based. The first was the appearance of a stereoscopic optic disc photograph as judged by masked experts; the second was the output of the Moorfields regression analysis (MRA) from the Heidelberg Retina Tomograph (HRT; Heidelberg Engineering, Dossenheim, Germany).

Methods

Subjects

One hundred eighty-five individuals with high-risk ocular hypertension or early glaucoma participated in this study, including 107 women and 78 men. The age range of this group was 37 to 87 years (mean ± SD: 60.9 ± 11.0). All subjects were fully informed with regard to the potential risks and benefits of the study and then provided voluntary written consent to participate. All procedures adhered to the tenets of the Declaration of Helsinki and were preapproved by the Legacy Health System Institutional Review Board for the protection of human subjects.

Participants were recruited prospectively from the Devers Eye Institute, or other ophthalmic practices in the Portland, Oregon, metropolitan area. All subjects had been tested on at least one previous occasion with full-threshold or SITA-standard white-on-white SAP (Carl Zeiss Meditec, Dublin, CA). At the time of recruitment, all subjects were considered to have either high-risk ocular hypertension or early glaucoma. Thus, all subjects had a history of untreated intraocular pressure ≥22 mm Hg in both eyes and at least one of the following additional risk factors: vertical cup-to-disc ratio ≥0.6 in at least one eye and/or an interocular cup-to-disc ratio asymmetry ≥0.2 between the two eyes; positive family history of glaucoma; personal history of migraine, Raynaud's syndrome, or vasospasm; African-American ancestry; or age >70 years. To be included, all subjects also met the following criteria for both eyes: best corrected visual acuity ≥20/40 and spectacle refraction < ±5.00 D sphere and < ±2.00 D cylinder. The definition of "high-risk" used for suspect eyes during the recruitment phase of this study might differ somewhat from that based on more recent studies, such as the Ocular Hypertension Treatment Study (OHTS).[21] Subjects were excluded if they had any other previous or current ocular or neurologic disease, previous ocular surgery (except uncomplicated cataract surgery), or diabetes requiring medication. SAP visual field status (normal or abnormal) was not a criterion for study entry, as long as the mean deviation (MD) was ≥ −6 dB at the time of recruitment.

Standard Automated Perimetry

SAP visual field testing was performed with the Humphrey Field Analyzer II (model 750; Carl Zeiss Meditec, Inc.). The 24-2 visual field (VF) test pattern was used for 119 (64%) of the subjects. The other 66 (36%) subjects were tested with the 30-2 pattern, but only the points corresponding to those in the 24-2 were analyzed. In nearly all tests, the SITA-standard threshold strategy was used (348/370 eyes; 94%), whereas the Full-Threshold strategy was used in the remaining 22 eyes. SAP visual field data were imported into a spreadsheet (Excel; Microsoft, Redmond, WA) for further analysis. The normative database for these analyses came from 348 eyes of 348 normal control subjects.[22] As previously described,[22] the data for each visual field test were transformed into 45-year-old equivalent sensitivities before calculation of visual field indices (MD, PSD, and GHT), and then compared to the results for normal 5%, 2%, 1%, and 0.5% probability levels for each of the global parameters, local threshold values and total deviation (TD) values.

SAP visual fields were considered abnormal if the pattern SD (PSD) value was beyond the normal 95% limit ($P \leq 0.05$) or if the Glaucoma Hemifield Test (GHT) was outside normal limits (OHTS criteria). Additional criteria for SAP VF abnormalities were also applied separately. Each of these criteria was based on a cluster of abnormal points. The definition of a cluster was two or three contiguous points within a single hemifield (i.e., completely contained within either the upper or the lower hemifield). Each of the various cluster criteria was specified as a unique combination of cluster size and probability level of the individual points within the cluster (Table 1). In this regard, these criteria were similar to those used for judging the mfVEP.

Multifocal VEP

Stimulus. Multifocal VEPs were obtained using one of the standard stimulus options available within the VERIS software package (Dart Board 60 with Pattern, VERIS 4; Electro-Diagnostic Imaging, San Mateo, CA). The stimulus was displayed on a 21-in. monochrome monitor (model VB 21300 GLS; Siemens AG, Munich, Germany) with a 75-Hz refresh rate. This stimulus has been used in previous studies of the mfVEP.[4,11,13,20,23] It has a total diameter of 44.5° and consists of 60 individual sectors (Fig. 1). Each sector contains 16 checks—8 white (200 cd/m^2) and 8 black (<1 cd/m^2)—and thus has a Michelson contrast of ~99%. The size of the individual stimulus sectors, as well as the size of individual checks, is scaled with eccentricity according to a cortical magnification factor. In all cases, the display was viewed through natural pupils with optimal refractive correction in place.

Table 1. Number and Percentage of Abnormal Eyes for Each of the SAP and mfVEP Cluster Definitions

Cluster Criterion	22	33	44	444	122	222	223	233	224	244	GHT/PSD (OHTS)
SAP											
Right eyes (*n*)	14	26	39	26	11	11	13	15	15	21	37
Left eyes (*n*)	29	40	57	42	20	20	26	27	28	32	46
% of 370	12%	18%	26%	18%	8%	8%	11%	11%	12%	14%	22%
mfVEP											
Right eyes (*n*)	34	45	74	44	16	17	24	28	27	36	
Left eyes (*n*)	54	64	93	70	35	35	41	48	51	66	
% of 370	24%	29%	45%	31%	14%	14%	18%	21%	21%	28%	
False alarm*	5%	13%	25%	11%	1%	1%	2.5%	3.5%	3%	8.5%	

There were 370 eyes (185 subjects). The numbers in the top row indicate the probability level of the points within the cluster (1 = $P < 0.005$; 2 = $P < 0.01$; 3 = $P < 0.02$; 4 = $P < 0.05$). Two numbers indicates that the cluster contains two contiguous points within a hemifield, and three numbers indicates three contiguous points within a hemifield. This analysis differs for SAP and mfVEP: for the mfVEP, an abnormality could be present on either the monocular SNR *or* interocular SNR ratio plot, whereas for SAP only monocular abnormalities were considered.

*False alarm rate taken from previously published results[23,24] from 100 control subjects.

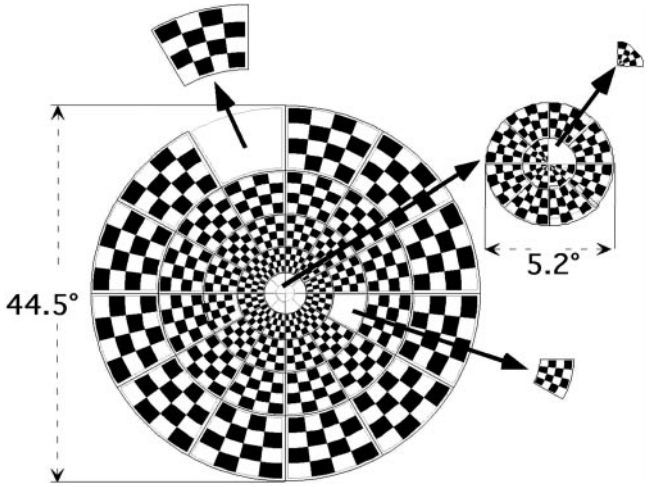

FIGURE 1. Schematic representation of mfVEP stimulus.

During a recording, the contrast polarity of each stimulus sector was temporally modulated (pattern reversals) according a pseudorandom m-sequence. The m-sequence chosen for this study had $2^{15} - 1$ steps in total and thus required ~8 minutes to complete. Each 8-minute recording was divided into 16 segments, and subjects were able to rest between segments as necessary. Each subject completed two recordings per eye, in an ABBA order, with the first eye chosen at random.

Recording Procedures. Three channels were recorded with gold disc electrodes (Grass model F-E5GH; Grass Technologies-Astro-Med, Inc., West Warwick, RI). For the vertical midline channel, electrodes were placed 4 cm above the inion (active) and at the inion (reference). For the left and right "oblique" channels, the active electrodes were placed 1 cm above the inion and 4 cm to the left or right, respectively, of the vertical midline; each of these active sites was referenced to the electrode at the inion. The left earlobe served as a ground for all three channels. In addition to the three channels recorded, data for three other channels were "derived" from the differences between various electrode pairs, as previously described.[11,16]

In preparation for recording, the skin at each electrode site was scrubbed with a gel preparation (Nuprep; D. O. Weaver & Co., Aurora, CO) on a cotton-tipped wooden swab. Electrodes were fixed in position with conductive cream (EC2; Astro-Med, Inc.) and secured with a polyester medical wrap (Coban; 3M, St. Paul, MN). Electrode impedance was maintained below 5 kΩ in all cases and was usually below 2 kΩ. Signals were amplified (Grass Model 12; Grass Technologies-Astro-Med, Inc.), band-pass filtered from 3 to 100 Hz (half-amplitude), and sampled at 1200 Hz (i.e., in 0.83-ms bins).

Signal-to-Noise Ratio. The two recordings for each eye were averaged after the mfVEP responses (the first slice of the second-order kernel) for each channel were exported from VERIS. No spatial smoothing or artifact rejection was applied before export. Additional mfVEP data analyses were performed with programs that were written in commercial software (MatLab; MathWorks Inc., Natick, MA). First, mfVEP records from each of the six channels were low-pass filtered by using a Fourier transform technique with a sharp cutoff at 35 Hz. Then, the signal-to-noise ratio (SNR) for each local response was derived as previously described.[11,18] The root-mean-square (RMS) amplitude of a signal window (45–150 ms) from each local response was divided by the RMS amplitude for a "noise window" (325–430 ms); the latter was taken as the *average* value of all 60 records for a given eye.[11,18]

The Best Array. All analyses were performed on the "best" of the responses from the six "channels" available for each eye.[11,16] These "best arrays", composed of the 60 "best SNR responses" for an individual, were derived differently for the monocular and interocular tests. In the monocular test, for each eye at each location, the response with the largest SNR among the six channels was selected for inclusion in the best array. For the interocular test, the response with the largest SNR was selected from the 12 responses (two eyes, six channels each eye) at each location. The response from the other eye from that same channel made up the pair of responses at that location in the interocular best array.

Abnormalities Defined by Cluster Criteria. The monocular SNR for each mfVEP response (i.e., the SNR at each location in each eye of each subject) was converted to a *z*-score relative to a previously published normal distribution derived from 100 control subjects.[24] Similarly, the interocular ratio (of root-mean-squared amplitude) at each location was converted to a *z*-score and assigned a probability value based on the corresponding local normal distribution. The mfVEP for each individual (i.e., the monocular mfVEP for the right and left eyes, as well as the interocular ratio) was determined to be either normal or abnormal based on various cluster criteria.[24,25] The definition of a cluster was two or three contiguous locations within a single hemifield. Each of the various cluster criteria were specified as a unique combination of cluster size and probability level (*z*-score) of the individual points within the cluster. The number of mfVEPs (eyes) classified as abnormal was compared for different cluster criteria, each associated with a known specificity.[24,25] The bottom row of Table 1 lists the estimate of false alarm rate (1 − specificity) corresponding to each cluster criterion, based on previous studies of control subjects with demographics similar to those of the present study population (100 individuals, age range: 22–92 years, average ± SD: 49.0 ± 13.6 years).

Stereo Optic Disc Photography and Grading

Optic disc photographs were obtained in all patients by using a simultaneous stereoscopic camera (3-Dx; Nidek Co., Ltd., Gamagori, Japan) after maximum pupil dilation. Two fellowship-trained glaucoma specialists (EP, AJ), masked to all other patient information, independently graded each photograph (viewed with a Stereo Viewer II; Asahi-Pentax, Tokyo, Japan) as either "normal" or "glaucomatous optic neuropathy" (GON) based on the following characteristics: adequate clarity and stereopsis, neuroretinal rim thinning (generalized or localized), excavation, retinal nerve fiber layer defect; violation of the "ISNT" rule, and cup-to-disc ratio. The presence and location of disc hemorrhages were noted, but in isolation, did not necessarily warrant a grade of GON. Disagreements between these two graders were adjudicated by a third masked expert (GAC).

Confocal Scanning Laser Ophthalmoscopy

Confocal scanning laser ophthalmoscopy (CSLO) images were obtained using the Heidelberg Retina Tomograph (versions 2.01/3.04 HRT; Heidelberg Engineering). Six 10° × 10° scans centered on the optic disc were acquired and the three judged to have the best quality were combined to create a mean topography for each eye. Experienced technicians outlined the margin of the optic disc while viewing stereo disc photographs. The overall MRA classification[26] was recorded for each eye. All eyes classified as "borderline" by the MRA were assigned to the normal category (i.e., "within normal limits"); this method maintains high specificity.[27] For example, in a separate group of 100 control subjects,[24] only 2 (1%) of 200 normal eyes were "outside normal limits," whereas 18 others were classified as "borderline" by the MRA.

Comparisons between SAP and mfVEP

SAP visual fields were obtained within 22.3 ± 27.0 days of the mfVEP. Stereo disc photographs and HRT images were obtained during the same visit, which was within 24.8 ± 50.4 days of the mfVEP and 33.1 ± 62.9 days of the SAP visual field. (Note the standard deviations are large because of a few outliers; the disc images were obtained within 3 months of the functional tests for 96% of subjects). Using the masked grade of each stereo disc photograph as the diagnostic classifier, sensitivity of the mfVEP or SAP was defined as the percentage of

photographs graded as GON that had an abnormality on the functional test. Specificity was defined as the percentage of photographs graded as normal that also had a normal functional test result. Similarly, when the HRT MRA result was used as the diagnostic classifier, sensitivity of the mfVEP or SAP was defined as the percentage of eyes ONL (on the HRT MRA) that also had an abnormality on the functional test. Specificity was defined as the percentage of eyes classified by the HRT MRA as within normal limits (including borderline eyes) that had a normal functional test result.

Results

Stereo Disc Photographs as a Diagnostic Standard

Disc photographs from 50% of eyes were graded GON (89 right eyes and 96 left eyes). Both eyes were graded GON in 71 (38%) subjects. Four photographs (both eyes of two subjects, 1%) could not be graded due to inadequate clarity and/or stereopsis; these four eyes were excluded from disc photo–based analyses. The average SAP MD was +0.3 ± 2.1 dB (range +3.9 to −10.1 dB) and the average PSD was 2.3 ± 1.9 dB (range, 1.0–16.1 dB). By OHTS criteria, 83 (22%) eyes had an abnormal SAP. By individual cluster criteria, the abnormality rate for SAP ranged from 8% to 26%, depending on the cluster criterion used. Table 1 shows the number (and percentage) of eyes with an abnormal SAP for each of the various cluster criteria evaluated. Only one of the individual cluster criteria (two points at $P < 0.005$; 26%) had a greater abnormality rate than the OHTS criterion.

The abnormality rate for mfVEP ranged from 14% to 45%, depending on the cluster criterion used. Table 1 shows the number (and percentage) of eyes with an abnormal mfVEP for each of the various cluster criteria evaluated. The cluster criteria were nominally identical with those used for SAP, except that for the mfVEP, an abnormality could be present on either the monocular SNR *or* interocular SNR ratio plot; whereas for SAP only monocular abnormalities were considered. The bottom row lists the estimate of false alarm rate (1 − specificity) corresponding to each cluster criterion, based on previous studies of control subjects with similar demographics.[24,25] For a cluster criterion with an estimated specificity of 91% (see "222," Table 1), 102 eyes (28%) had an abnormal mfVEP. For cluster criteria with specificities of 95% ("22") and 99% ("222"), 88 (24%) eyes and 52 (14%) eyes, respectively, had an abnormal mfVEP. Agreement between the mfVEP and SAP (OHTS criterion) was generally close to 80%, ranging from 75% to 81%, depending on the cluster criterion used for mfVEP classification.

When the masked grade of each stereo disc photograph was used for diagnostic classification, 54 of the 185 eyes graded as GON had an abnormal SAP (by OHTS criteria), whereas 152 of the 181 discs graded as normal had a normal SAP. Thus, the sensitivity (95% CI) of SAP using OHTS criteria to detect GON was 29% (24%–34%), whereas specificity (95% CI) was 84% (80%–88). These results for SAP are shown by the filled symbols in Figure 2A (for right and left eyes, separately). Figure 2A also shows the results for each of the various SAP cluster criteria (open symbols). Compared with the OHTS criterion, these cluster criteria generally had lower sensitivity, but higher specificity for detecting eyes graded as GON by stereo disc photographs. Considering all eyes together, only one SAP cluster criterion had both higher sensitivity and specificity than the OHTS criterion: the "44" criterion (i.e., two points at $P < 0.05$; Table 1) had a sensitivity of 35% (30%–41%) and specificity of 85% (79%–90%).

The results of the same analysis for the mfVEP are shown in Figure 2B, where sensitivity of the mfVEP to detect eyes graded as GON is plotted against specificity for each of the cluster criteria shown in Table 1. To enable comparison between SAP and mfVEP, the OHTS criterion results from Figure 2A are plotted in Figure 2B (filled symbols). Further, the range of mfVEP criteria can be narrowed to only those with specificities similar to SAP. For example, the sensitivity of the mfVEP to detect GON, for cluster criteria with specificity between 84% and 87%, ranged from 28% to 32%. The mfVEP cluster criterion providing the specificity most similar to that of the SAP-OHTS criterion was the "22" type (i.e., two points at $P < 0.01$; Table 1), which had a sensitivity of 32% (27%–37%) and specificity of 84% (79%–89%) for detection of disc photograph graded GON. This performance is statistically indistinguishable from that of SAP, when using either the SAP-OHTS criterion or the SAP "44" cluster criterion. Thus, the results of the two functional tests are very similar when disc photograph–graded GON is used as the diagnostic standard.

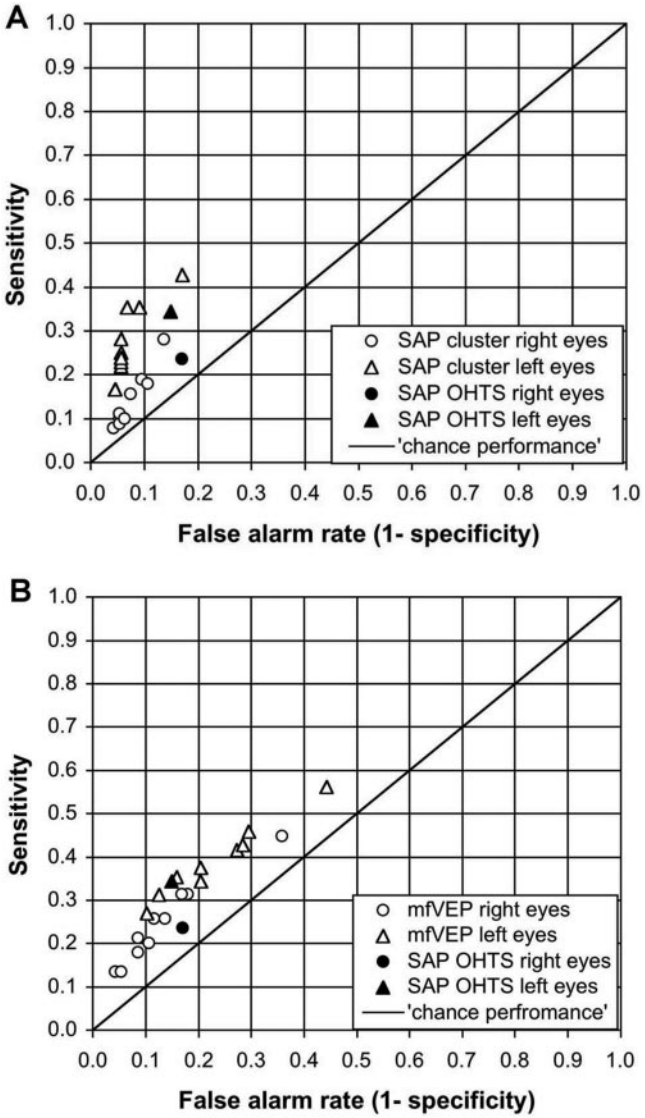

Figure 2. Sensitivity versus false alarm rate for SAP when masked grades of stereo disc photographs were used as the diagnostic standard (**A**). Abnormalities on SAP were determined using either the OHTS criterion or various clusters. Sensitivity versus false-alarm rate for mfVEP when masked grades of stereo disc photographs are used as the diagnostic standard (**B**). The data for SAP (OHTS criterion) are replicated from (**A**) for reference.

HRT MRA as a Diagnostic Standard

The MRA classification from the HRT was ONL in 25% of the eyes (46 right eyes and 47 left eyes). Both eyes were ONL in 30 (16.4%) subjects. The HRT scan had insufficient quality in four (1%) eyes; these eyes were excluded from HRT-based analyses. The MRA classified 63 (17%) eyes as "borderline." Agreement between masked grades of stereo disc photographs and the MRA classifications was 76% when the borderline cases were assigned to the GON category. Agreement was 69% when the borderline MRA cases were assigned to the normal category. For this study, borderline MRA cases were assigned to the normal category, to maintain higher specificity.[27] Of the 252 eyes where the two structural classifications agreed, 82 (33%) eyes were determined to have GON, thus 170 (67%) were classified as normal according to both the disc photographs and the MRA.

When the MRA result from the HRT was used for diagnostic classification, 39 of the 93 eyes that were ONL had an abnormal SAP (by OHTS criteria), whereas 231 of the 273 discs that were within normal limits (including borderline) had a normal SAP. Thus, sensitivity of SAP-OHTS to detect GON as determined by the HRT MRA was 42% (37%–48%), while specificity was 85% (80%–91%). These data are plotted in Figure 3 (filled diamond), which also shows the family of SAP cluster results (open circles) using the HRT MRA as the diagnostic standard. Figure 3 shows again that the SAP-OHTS criterion generally had higher sensitivity, but lower specificity, than most of the SAP cluster criteria. Of the SAP cluster types evaluated, only the "44" (two points at $P < 0.05$) and "444" (three points at $P < 0.05$) performed better than SAP-OHTS. The "44" type had higher sensitivity (54%; 47%–61%) with similar specificity (85%), whereas the "444" type had similar sensitivity (44%) but higher specificity (91%; 85%–97%) than SAP-OHTS.

Figure 3 also shows the results for the mfVEP (open triangles) when the HRT MRA classification was used as the diagnostic standard. The results of the two functional tests are very similar. For cluster criteria with specificities between 83% and 91%, the sensitivity of the mfVEP to detect GON, as determined by the HRT MRA, ranged from 29% to 45%. The cluster criterion with specificity closest to the performance of SAP-OHTS, namely a three-point cluster ($P < 0.01, 0.01, 0.05$; "224" in Table 1) within a single hemifield on *either* the monocular SNR *or* the interocular plot, had a sensitivity to detect GON of 42% (36%–48%) and a specificity of 86% (80%–92%), essentially identical with the values found for SAP-OHTS and just slightly lower than the sensitivity of SAP "44" or the specificity of SAP "444." The mfVEP cluster type "22" ($P < 0.01, 0.01$) provided just slightly higher sensitivity (45%) but slightly lower specificity (83%).

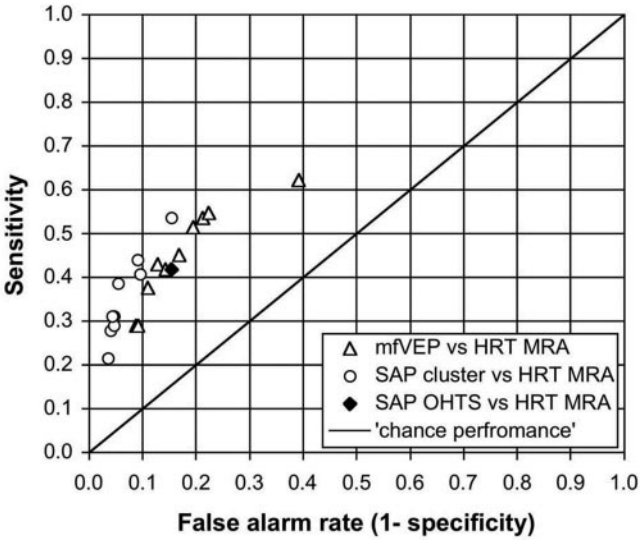

FIGURE 3. Sensitivity versus false-alarm rate for mfVEP and SAP when the HRT MRA was used as the diagnostic standard. Abnormalities on SAP were determined using either the OHTS criterion or various cluster criteria. Abnormalities on mfVEP were determined according to various cluster criteria. Note that right and left eyes were combined for presentation in this graph.

Stereo Disc Photographs and the HRT MRA Combined as a Diagnostic Standard

In addition to the comparisons of sensitivity and specificity, an analysis of disagreements between the two functional tests was also performed. When SAP and mfVEP agree, there is no differential effect on their relative sensitivity–specificity when one or another imperfect diagnostic standard is used. However, given that disagreements between SAP and mfVEP could be informative in terms of their relative performance, a "combined" structural diagnostic standard was applied because it provides a more definitive index against which to compare the two functional tests.

Using this combined diagnostic standard, there were 252 eyes in which the disc photograph grade and HRT MRA agreed: 170 eyes with normal optic disc structure and 82 eyes with glaucomatous optic discs (GON). SAP (by OHTS) and the combined disc classification were both normal in 145 (58%) eyes and abnormal in 37 (35%) eyes. SAP was normal in 45 of 82 eyes that were called GON by the combined disc classification (55% "miss rate"). SAP was abnormal in 25 of the 170 eyes that were called normal by the combined disc classification (15% "false alarm rate"). Thus, the sensitivity of SAP-OHTS to detect GON (per combined diagnostic standard) was 45% (39%–51%) and the specificity was 85% (79%–91%). Again, the performance of two SAP cluster criteria exceeded that of SAP-OHTS for one dimension each: a two point cluster at $P < 0.05$ ("44") provided slightly greater sensitivity (56%; 49%–63%) and a three-point cluster at $P < 0.05$ ("444") provided slightly greater specificity (92%; 85%–99%) than SAP-OHTS.

The percentage of these 252 eyes that were classified as normal by both the combined disc classification and the mfVEP ranged from 41% to 63% (average, 56%) depending on the mfVEP cluster criterion used. The percentage of eyes that were called abnormal by both mfVEP and the combined disc classification ranged from 10% to 21% (average, 15%). Of the 82 eyes that were called GON by the combined disc classification, the mfVEP was normal in 29 to 56 (35%–68% "miss rate"; average, 53%). The mfVEP was abnormal in 12 to 66 of the 170 eyes that were called normal by the combined disc classification (7% to 39% "false alarm rate," average, 16%). Thus, the average performance of mfVEP was essentially identical with SAP-OHTS. The three best mfVEP cluster criteria had approximately the same performance: The "22" ($P < 0.05, 0.05$), the "233" ($P < 0.01, 0.02, 0.02$) and the "224" ($P < 0.01, 0.01, 0.05$) had sensitivities between 43% and 46% and specificities between 85% and 88%—that is, equivalent to SAP-OHTS and slightly inferior to the two best SAP cluster criteria (although 95% CIs were all approximately ±5% and thus overlapped).

Although the mfVEP and SAP again had a similar overall performance when judged against the combined structural diagnostic standard, there were still a substantial number of individual cases in which the two functional tests disagreed, which was the primary focus of this analysis. Table 2 shows that disagreement between the two functional tests ranged from 51 to 85 of these 252 eyes (20%–34%, generally closer to 20%), depending on which mfVEP cluster criterion was com-

TABLE 2. Comparison between SAP (OHTS criterion) and mfVEP When the Two Functional Tests Did Not Agree

mfVEP Cluster Criterion	22	33	44	444	122	222	223	233	224	244
GON disc	12/23	16/24	20/24	15/21	7/25	7/25	10/24	12/23	10/22	14/22
	52%	67%	83%	71%	28%	28%	42%	52%	45%	64%
Normal disc	17/34	23/36	51/61	25/40	7/27	7/26	11/30	13/30	16/35	22/37
	50%	64%	84%	63%	26%	27%	37%	43%	46%	59%

The top row shows disagreement when the diagnostic classification (based on combined HRT MRA and stereo disc photo grade) was GON, and the bottom row shows disagreement among eyes with normal optic disc structure. Each cell in the table shows the number (*n*) of eyes with an abnormal mfVEP and a normal SAP along with the total number (T) of eyes in which they disagreed as *n*/T with the percentage given below these two numbers.

pared with the SAP-OHTS criterion. Each cell in Table 2 lists the number (N) of eyes with an abnormal mfVEP and a normal SAP (by OHTS) along with the total number (T) of eyes in which they disagreed as N/T, with the resultant percentage given below the two numbers. Table 2 shows that there were just over 20 cases of GON and approximately 30 to 40 eyes with normal optic disc structure (by the combined standard), in which the two functional tests disagreed. Thus, the likelihood of disagreement between the two functional tests was ~50% greater when optic disc structure was normal rather than glaucomatous.

Table 2 also shows that in approximately half of the cases where the two functional tests disagree, the mfVEP agrees with the diagnostic standard (optic disc structural classification), whereas in the other half, SAP agrees with the diagnostic standard. For example, evaluating mfVEP cluster criteria with the most similar overall performance to SAP-OHTS (e.g., "233" or "224" see the Results section) reveals that when the mfVEP and SAP disagree, they are about equally likely to agree (or disagree) with the optic disc structural classification. Figure 4 presents the results of comparing disagreements between SAP-OHTS and mfVEP ("224" criterion) among the group of eyes limited to the combined structural diagnostic standard. Figure 4 shows that SAP and mfVEP agreed in 77% of these 252 eyes and the majority (61%) of their disagreements were in eyes with normal optic disc structure. When SAP and mfVEP disagreed, they were equally likely to agree (or disagree) with the structural diagnostic standard.

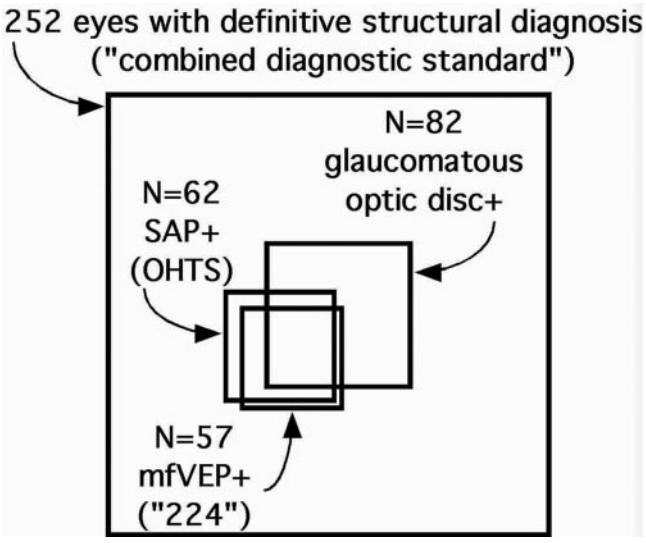

FIGURE 4. Agreement between SAP (OHTS criterion), mfVEP ("224" cluster criterion) and the combined diagnostic standard (disc photograph grade plus HRT MRA). The area–proportional Venn diagram shows the number of eyes in which the disc photograph grade and HRT MRA classification matched (outer box, $n = 252$), the number and proportion of those eyes classified as having a glaucomatous optic disc ($n = 82$), the number and proportion with an abnormal SAP ($n = 62$), and the number and proportion with an abnormal mfVEP ($n = 57$). Additional numerical data are available in Table 2, 224 cluster.

DISCUSSION

The results of this study demonstrate that the diagnostic performance of mfVEP was similar to that of SAP. The sensitivity of these two functional tests to detect GON was equal when matched for specificity, regardless of whether the diagnostic standard was based on masked experts' grading of a stereo disc photo, the HRT MRA, or the combination of both structural assessments. These findings are consistent with those of Graham et al.[14] who used disc photographs as an independent diagnostic standard. In our study, the sensitivity to detect GON was higher for both functional tests (42%) when the HRT MRA was used as the diagnostic standard, compared with when the disc photograph grade was used as the diagnostic standard (29%). This result is explained by the fact that the HRT MRA was generally more conservative than the masked expert graders, classifying exactly half as many eyes as having GON. The specificity of both functional tests remained relatively high (85%, i.e., the false alarm rate did not increase when the HRT MRA was used in place of disc photograph grades), indicating that nearly all the eyes in which the two structural classifiers disagreed had normal results on both functional tests. This suggests that there was better agreement between both functional tests and the HRT MRA than with the stereo disc photograph grades.

Agreement between the two functional tests was also generally good (~80%). This is consistent with the findings of Bjerre et al.[12] who reported "fair" agreement between SAP and mfVEP in subjects diagnosed with (manifest) glaucoma. Perhaps the more interesting portion of the population is the 20% of eyes in which the two functional tests did not agree, notwithstanding that the criteria used to define abnormalities, as well as the normative databases differed. Hood et al.[11,13,20] predicted that the performance of a monocular mfVEP test would be approximately equal to that of SAP, but slightly better than SAP when the interocular analysis was added to the mfVEP. We used an "either-or" combination of monocular and interocular tests to define an abnormality for the mfVEP, but restricted the definition of an abnormal SAP to only a monocular analysis. The proportion of eyes with an abnormal SAP, as well as the agreement with mfVEP might have been even greater if point-wise interocular analyses had also been used for SAP.[28] Further studies are needed to establish normative ranges and evaluate potential specificity tradeoffs for point-wise interocular SAP threshold asymmetry analyses.

The fact that SAP and mfVEP performed equally well, but agreed in only ~80% of cases, suggests that the mfVEP detects

some real abnormalities that SAP is missing and vice versa. There are also differences between the spatial patterns and contiguity of test points. The current body of evidence (see e.g., work of Hood et al.[11,13,20]) indicates that in cases where the mfVEP SNR is high, it is more likely than SAP to detect functional abnormalities using the point-wise interocular comparison, whereas in cases where the mfVEP SNR is low, the interocular comparison becomes more variable and the mfVEP advantage is lost (i.e., SAP will be more sensitive for a given specificity). The mfVEP is more likely than SAP to miss defects in the superior periphery[13] because SNR is generally lower at those stimulus locations,[24] whereas SAP is more likely to miss localized central defects[13] because of its lower spatial resolution at those locations. It is also likely that some portion of the disagreement is attributable to variability (noise). For example, in a population with similar characteristics, >85% of initial defects on SAP were not confirmed on retest.[29] Recent work suggests that mfVEP abnormalities (clusters) rarely repeat in healthy control eyes, especially in the same location.[25] Thus, confirmation of a mfVEP cluster abnormality in an early glaucoma or suspect eye is likely to represent a real (reliable) functional defect. Longitudinal evaluation will help determine which of the functional abnormalities in these high-risk patients with suspected or early glaucoma are repeatable and associated with progressive GON.

Agreement between the two *structural* classifiers was lower, 69% or 76%, depending on whether borderline MRA cases were assigned to the normal or GON category. We chose to assign the borderline MRA cases to the normal category, to maintain higher specificity.[27] In fact, the number of eyes classified as GON by the HRT MRA increased by 60% when the borderline cases were switched to the GON category. This finding underscores the fact that many of the subjects were referred in to the study as high risk for suspected glaucoma because they had a suspicious (or borderline) optic disc appearance, yet had a normal or nearly normal visual field in one or both eyes and thus may represent a bias in our study population that would lower the apparent sensitivity of both functional tests. Previous studies have shown that the proportion of eyes with functional abnormalities (and the positive predictive value of such tests for GON) is relatively low, even in eyes with cup-to-disc ratios ≥0.8.[30,31] It should also be noted that as the definition of high-risk suspect eyes continues to evolve, it might differ from that used during the recruitment phase of this study, although this should not have a major impact on the results of this cross-sectional comparison of SAP and mfVEP.

In summary, the sensitivity and specificity of mfVEP and SAP were similar when the diagnostic standard was based on structural characteristics of the optic disc. Agreement between the two functional tests was ~80%, which was higher than the agreement between the two measures of optic disc structure. Although the specificity of both functional tests was reasonably high (~85%), sensitivity to detect GON was relatively low (~30%–45%). The low sensitivity may be partially attributable to the selection bias of the study population, but also suggests that structural and functional abnormalities are not highly coincident during early-stage glaucoma.[32] This notion is supported by the results from OHTS in which approximately 60% of all conversions occurred because of structural progression (optic disc change), 40% because of changes on SAP, but <15% by changes in both structure and function.[21] Thus, the absolute values of sensitivity and specificity derived from the present study should be interpreted with caution, as they depend greatly on the diagnostic standard applied, the composition of the study population, and the nature of structure–function relationships in early glaucoma.

Acknowledgments

The authors thank Cindy Blachly, Thie Smith, Judy Thompson, and Karin Novitsky for their care and diligence during data collection.

References

1. Klistorner AI, Graham SL, Grigg JR, Billson FA. Multifocal topographic visual evoked potential: improving objective detection of local visual field defects. *Invest Ophthalmol Vis Sci*. 1998;39:937–950.
2. Klistorner A, Graham SL. Objective perimetry in glaucoma. *Ophthalmology*. 2000;107:2283–2299.
3. Graham SL, Klistorner AI, Grigg JR, Billson FA. Objective VEP perimetry in glaucoma: asymmetry analysis to identify early deficits. *J Glaucoma*. 2000;9:10–19.
4. Hood DC, Zhang X, Greenstein VC, et al. An interocular comparison of the multifocal VEP: a possible technique for detecting local damage to the optic nerve. *Invest Ophthalmol Vis Sci*. 2000;41:1580–1587.
5. Hood DC, Zhang X. Multifocal ERG and VEP responses and visual fields: comparing disease-related changes. *Doc Ophthalmol*. 2000;100:115–137.
6. Hasegawa S, Abe H. Mapping of glaucomatous visual field defects by multifocal VEPs. *Invest Ophthalmol Vis Sci*. 2001;42:3341–3348.
7. Bengtsson B. Evaluation of VEP perimetry in normal subjects and glaucoma patients. *Acta Ophthalmol Scand*. 2002;80:620–626.
8. Goldberg I, Graham SL, Klistorner AI. Multifocal objective perimetry in the detection of glaucomatous field loss. *Am J Ophthalmol*. 2002;133:29–39.
9. Fortune B, Goh K, Demirel S, et al. Detection of glaucomatous visual field loss using the multifocal visual evoked potential. In: Henson DB, Wall M, eds. *Perimetry Update 2002/2003; Proceedings of the XVth International Perimetric Society Meeting.* The Hague: Kugler; 2003;251–260.
10. Thienprasiddhi P, Greenstein VC, Chen CS, et al. Multifocal visual evoked potential responses in glaucoma patients with unilateral hemifield defects. *Am J Ophthalmol*. 2003;136:34–40.
11. Hood DC, Greenstein VC. Multifocal VEP and ganglion cell damage: applications and limitations for the study of glaucoma. *Prog Retin Eye Res*. 2003;22:201–251.
12. Bjerre A, Grigg JR, Parry NR, Henson DB. Test-retest variability of multifocal visual evoked potential and SITA standard perimetry in glaucoma. *Invest Ophthalmol Vis Sci*. 2004;45:4035–4040.
13. Hood DC, Thienprasiddhi P, Greenstein VC, et al. Detecting early to mild glaucomatous damage: a comparison of the multifocal VEP and automated perimetry. *Invest Ophthalmol Vis Sci*. 2004;45:492–498.
14. Graham SL, Klistorner AI, Goldberg I. Clinical application of objective perimetry using multifocal visual evoked potentials in glaucoma practice. *Arch Ophthalmol*. 2005;123:729–739.
15. Baseler HA, Sutter EE, Klein SA, Carney T. The topography of visual evoked response properties across the visual field. *Electroencephalogr Clin Neurophysiol*. 1994;90:65–81.
16. Hood DC, Zhang X, Hong JE, Chen CS. Quantifying the benefits of additional channels of multifocal VEP recording. *Doc Ophthalmol*. 2002;104:303–320.
17. Klistorner AI, Graham SL. Electroencephalogram-based scaling of multifocal visual evoked potentials: effect on intersubject amplitude variability. *Invest Ophthalmol Vis Sci*. 2001;42:2145–2152.
18. Zhang X, Hood DC, Chen CS, Hong JE. A signal-to-noise analysis of multifocal VEP responses: an objective definition for poor records. *Doc Ophthalmol*. 2002;104:287–302.
19. Bowd C, Zangwill LM, Berry CC, et al. Detecting early glaucoma by assessment of retinal nerve fiber layer thickness and visual function. *Invest Ophthalmol Vis Sci*. 2001;42:1993–2003.
20. Hood DC, Zhang X, Winn BJ. Detecting glaucomatous damage with multifocal visual evoked potentials: how can a monocular test work? *J Glaucoma*. 2003;12:3–15.

21. Gordon MO, Beiser JA, Brandt JD, et al. The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. *Arch Ophthalmol.* 2002;120:714–720; discussion 829–730.
22. Johnson CA, Sample PA, Cioffi GA, Liebmann JR, Weinreb RN. Structure and function evaluation (SAFE): I. criteria for glaucomatous visual field loss using standard automated perimetry (SAP) and short wavelength automated perimetry (SWAP). *Am J Ophthalmol.* 2002;134:177–185.
23. Fortune B, Hood DC. Conventional pattern-reversal VEPs are not equivalent to summed multifocal VEPs. *Invest Ophthalmol Vis Sci.* 2003;44:1364–1375.
24. Fortune B, Zhang X, Hood DC, Demirel S, Johnson CA. Normative ranges and specificity of the multifocal VEP. *Doc Ophthalmol.* 2004;109:87–100.
25. Fortune B, Demirel S, Zhang X, Hood DC, Johnson CA. Repeatability of normal multifocal VEP: implications for detecting progression. *J Glaucoma.* 2006;15:131–141.
26. Wollstein G, Garway-Heath DF, Hitchings RA. Identification of early glaucoma cases with the scanning laser ophthalmoscope. *Ophthalmology.* 1998;105:1557–1563.
27. Ford BA, Artes PH, McCormick TA, et al. Comparison of data analysis tools for detection of glaucoma with the Heidelberg Retina Tomograph. *Ophthalmology.* 2003;110:1145–1150.
28. Levine R, Demirel S, Fan J, et al. The Ocular Hypertension Treatment Study Group. Asymmetries and visual field summaries as predictors of glaucoma in the ocular hypertension treatment study (OHTS). *Invest Ophthalmol Vis Sci.* 2006;47:3896–3903.
29. Keltner JL, Johnson CA, Quigg JM, et al. Confirmation of visual field abnormalities in the Ocular Hypertension Treatment Study. Ocular Hypertension Treatment Study Group. *Arch Ophthalmol.* 2000; 118:1187–1194.
30. Mansberger SL, Sample PA, Zangwill L, Weinreb RN. Achromatic and short-wavelength automated perimetry in patients with glaucomatous large cups. *Arch Ophthalmol.* 1999;117:1473–1477.
31. Mansberger SL, Johnson CA, Cioffi GA, et al. Predictive value of frequency doubling technology perimetry for detecting glaucoma in a developing country. *J Glaucoma.* 2005;14:128–134.
32. Keltner JL, Johnson CA, Anderson DR, et al. The association between glaucomatous visual fields and optic nerve head features in the Ocular Hypertension Treatment Study. *Ophthalmology.* 2006; 113:1603–1612.