

# Evaluating Several Sources of Variability for Standard and SWAP Visual Fields in Glaucoma Patients, Suspects, and Normals

Eytan Z. Blumenthal, MD,<sup>1</sup> Pamela A. Sample, PhD,<sup>1</sup> Charles C. Berry, PhD,<sup>2</sup> Alexander C. Lee, BA,<sup>1</sup> Christopher A. Girkin, MD,<sup>1</sup> Linda Zangwill, PhD,<sup>1</sup> Joseph Caprioli, MD,<sup>3</sup> Robert N. Weinreb, MD<sup>1</sup>

**Purpose:** To quantify factors affecting test-retest variability of threshold measurements over a series of 3 serial visual fields (VF).

**Design:** Prospective comparative observational study.

**Participants:** Forty-one normals, 10 suspects and 35 stable glaucoma patients.

**Methods:** All subjects performed 3 standard and 3 short-wavelength automated perimetry (SWAP) VFs. At each VF location, severity (defined as age-corrected total deviation) and test-retest variability (TRV), defined as the standard deviation of 3 serial threshold values, were calculated. A multiple regression model (constructed separately for standard VF and SWAP) incorporated 13 factors: severity, location, eccentricity, study group, diagnosis, superior versus inferior hemifield, nasal versus temporal hemifield, one-versus-two thresholds, age, mean pupil size, pupil size variability, between-subject variation, and residual variation.

**Main Outcome Measures:** Variability in threshold sensitivity VF values.

**Results:** Mean TRV ( $\pm$  standard deviation) for normal, suspect and glaucoma eyes, respectively, was:  $1.28 \pm 0.87$ ,  $1.53 \pm 1.04$  and  $2.20 \pm 1.79$  dB for standard VF, and  $1.87 \pm 1.35$ ,  $1.86 \pm 1.24$  and  $2.68 \pm 1.85$  dB for SWAP. The contribution of each factor to the model for standard VF and SWAP (SWAP in parentheses) were: severity 15.5% (6.9%); location 2.7% (4.1%); eccentricity 1.1% (0.64%); diagnosis 2.9% (5.9%); "superior versus inferior" hemifield 0.17% (1.7%); "nasal versus temporal" hemifield 0.06% (0.02%); one-versus-two thresholds 0.04% (0.16%); age 0.1% (0.06%); mean pupil size 0.59% (0.1%); pupil size variability 3.2% (2.8%); between-subject 8.0% (13.5%) and residual variation 61.0% (66.6%). Excluding between-subject and residual variation, the 11-factor model was able to account for less than one third of the variability seen in both standard VF and SWAP.

**Conclusions:** Severity of defect and between subject variation exerted the largest effect on TRV. However, even if all 11 factors could be adjusted for, it would reduce the magnitude of TRV by only 30%. More work is needed to reduce the remaining variability inherent in psychophysical testing and to better understand the intrinsic physiological variability present both in healthy and diseased eyes. It is possible that a larger number of VFs used for the calculation of TRV might further reduce the magnitude of the remaining variability found in this study. *Ophthalmology* 2003;110:1895–1902 © 2003 by the American Academy of Ophthalmology.

Originally received: July 24, 2002.

Accepted: December 5, 2002.

Manuscript no. 220502.

<sup>1</sup> Glaucoma Center and Visual Function Laboratory, Dept. of Ophthalmology, University of California, San Diego, California.

<sup>2</sup> Dept. of Family and Preventive Medicine, University of California, San Diego, California.

<sup>3</sup> Dept. of Ophthalmology, University of California, Los Angeles, California.

Presented in part at the Association for Research in Vision and Ophthalmology meeting, Ft Lauderdale, Florida, May 2000.

The authors have no financial interest in the perimetry technology presented. Supported in part by NIH grants EY08208 (PAS), EY11008 (LZ), EY12738 (JC), Bethesda, Maryland; Glaucoma Research Foundation (PAS), San Francisco, California; Research to Prevent Blindness, Lew R. Wasserman Award (PAS and JC), New York, New York; Joseph Drown Foundation (RNW), Los Angeles, California; and Foundation for Eye Research (EZB), Rancho Santa Fe, California. EZB is a recipient of a Fellowship from the American Physicians Fellowship for Medicine in Israel, Jerusalem, Israel.

Correspondence to Pamela Sample, PhD, University of California, San Diego, Dept of Ophthalmology-0946, 9500 Gilman Drive, La Jolla, CA 92093-0946. E-mail: psample@eyecenter.ucsd.edu.

Visual fields are the accepted standard for determining and documenting the status of the visual function in glaucoma.<sup>1</sup> Likewise, visual fields are an accepted method for judging progression of glaucomatous optic neuropathy. Progression is assessed by detecting change in visual field (VF) threshold measurements.

For analysis of change, one must consider data quantifying the variability of VF threshold measurements,<sup>2</sup> and this, in turn, must rely on understanding the various sources underlying such variability. One example is the shift in prevailing practice patterns from VF testing using a Humphrey VF analyzer 30-2 to a 24-2 testing grid in both clinical practice and research settings. Besides shortening the test duration, this change in practice occurred, at least in part, due to the larger variability found in the peripheral ("edge") points of the 30-2 grid.

Several factors have previously been shown to affect test-retest variability (TRV) in VFs. Several studies established that depressed areas of the visual field, manifesting as locations with reduced threshold sensitivities, are accompanied by an increased TRV.<sup>3-6</sup> Flammer suggested that an increased TRV might itself be an indication of glaucoma.<sup>7</sup> Similarly, greater eccentricity of a test point location is accompanied by higher variability of threshold measurements.<sup>3-5,8</sup> Aging, while associated with reduced thresholds,<sup>9</sup> is conversely associated with increased long- as well as short-term variability.<sup>10</sup> Pupil size, for normal subjects, was shown to influence threshold values both when constricted<sup>11</sup> and when dilated.<sup>12</sup> When glaucoma patients with constricted pupils secondary to chronic pilocarpine therapy were dilated, their mean defect improved an average of 3.1 dB.<sup>13</sup> Likewise, short-wavelength automated perimetry (SWAP), aimed at testing only a subset of ganglion cells, was shown to be more variable than standard VF for normals<sup>14</sup> and glaucoma subjects.<sup>15</sup>

It is, however, unclear which factors exert clinically meaningful and independent effects on TRV. In addition, little is known about the relative magnitude of effect certain factors have on TRV. To address this issue, a model incorporating the relevant factors was developed. Such a model is necessary to estimate the independent contribution of each factor on variability of threshold measurements.

Two examples highlight potential conclusions that may be drawn from this analysis. First, is eccentricity independently related to TRV or does it interact with severity to influence TRV? Since glaucoma damage is often more evident in peripheral locations on the 24-2 grid, it is possible that severity interacts with eccentricity to cause the higher TRV. Second, do glaucomatous eyes have a tendency towards increased TRV compared with normal eyes, or is increased TRV merely a result of more severe defects?

Estimating TRV is important for determining progression of VF loss. For progression to be reliably determined it is important to first determine, on a point-by-point basis, the magnitude of TRV that is not associated with glaucomatous progression. Any change greater than this TRV might then more likely indicate progression. While this study aims at studying variability of individual VF locations, it has been shown that combining locations in nerve fiber bundle patterns can further improve our ability to determine progression.<sup>16</sup>

In this study the individual effects, as well as the combined effect, of 13 factors on TRV seen in serial VF threshold measurements were estimated. Our approach was to present factors related to TRV in a way that may later benefit those designing algorithms for detection of progression in individual patients.

## Patients and Methods

### Patients

Eighty-six subjects (41 normals, 10 suspects and 35 glaucoma patients) were included in the study. Prior to the study, all patients had a complete ophthalmologic examination, including visual acuity, slit-lamp exam, gonioscopy, applanation tonometry, and a

dilated fundus exam. Informed consent was obtained from all participants and the University of California, San Diego Human Subject Committee approved all methodology.

Exclusion criteria on baseline exam included: field loss threatening fixation in either eye; history of acute angle closure, congenital glaucoma, secondary glaucoma, or ocular trauma; history of ocular infection or inflammatory disease within the past 6 months; narrow angles or other angle abnormalities, previous intraocular surgery; use of systemic medications that may affect intraocular pressure; use of miotic medications; history of severe retinal disease; best-corrected distance visual acuity worse than 20/30; strabismus; refractive error greater than  $\pm 5.00$  diopters spherical equivalent, or greater than  $\pm 3.00$  diopters of cylinder; lens opacity exceeding Lens Opacity Classification System (LOCS) III standard photographs,<sup>17</sup> nuclear color grade 4, nuclear opalescence grade 4, cortical cataract grade 3 or posterior subcapsular cataract grade 2 in either eye; pupil size  $< 3$  mm; or congenital color vision defects. Patients who had undergone medication changes during the study period or with a history of color vision problems or any other diseases known to affect visual function, such as diabetes or age-related macular degeneration were excluded. None of the subjects underwent any ocular surgery during the study period. The normal II group was not evaluated using the LOCS III classification; however, for this group the inclusion criteria were visual acuity of 20/25 or better and clear media apart from sclerotic changes in the lens.<sup>14</sup>

Glaucoma patients had intraocular pressures  $\geq 23$  mmHg on at least 2 separate visits, a glaucomatous optic disc and repeatable VF loss. Glaucomatous optic disc was defined as having evidence of rim thinning, notching, excavation, cup-disc asymmetry between the two eyes of greater than 0.2, or characteristic nerve fiber layer defects, as determined by masked stereo-photograph grading. An abnormal VF required a CPSD outside 95% normal limits or a glaucoma hemifield test "outside normal limits." Patients with advanced loss (defined as a mean deviation [MD] worse than  $-15$  dB on the standard VF), or VF progression during the study period, were excluded. A formal progression analysis is described below.

Glaucoma suspects had normal VFs, an intraocular pressure greater than 23 mmHg on at least 2 separate occasions and/or glaucomatous-appearing optic disks, as described above. Normal patients had intraocular pressures of less than 21 mmHg (and no history of elevated intraocular pressure), normal-appearing optic discs, normal visual fields, and no family history of glaucoma. The Normal and Glaucoma subjects, each, were composed of 2 subgroups based on the time interval allowed between the 3 VF tests (in the case of Glaucoma I & II) and the recruiting sites (in the case of Normal I and II). This subgroup division is accounted for in the statistical modeling as "study groups." Demographic differences between the subgroups are listed in Table 1.

### Visual Field Testing

All subjects performed at least 2 standard VFs prior to the commencement of the study. This past experience was required to reduce the learning curve effect. To test for any remaining residual learning curve effect, a formal analysis, incorporating MD values, was performed.

All patients in the normal I, suspect, and Glaucoma I groups performed 3 standard (white-on-white) and 3 SWAP (blue-on-yellow) VFs, separated by 1 week and completed within 2 weeks. All testing, both standard automated perimetry and SWAP, was performed using the 24-2 full threshold (using the 4-2-2 threshold strategy) program of the commercially available Humphrey Field Analyzer (Humphrey-Zeiss, Dublin, California). Both the background and stimulus intensities of all HFAI and II units used were calibrated prior to the initiation of the study. Both standard and

Table 1. Patient Demographics

	Normals I (n = 10)	Normals II (n = 31)	Suspects (n = 10)	Glaucoma I (n = 10)	Glaucoma II (n = 25)
Age: mean $\pm$ SD days (range)	64.4 $\pm$ 3.7 (59–70)	35.5 $\pm$ 6.6 (26–49)	55.7 $\pm$ 11.5 (35–70)	62.2 $\pm$ 12.2 (44–80)	61.0 $\pm$ 12.3 (39–79)
Eye: RE/LE	4/6	NA	3/7	4/6	13/12
Interval: 1 <sup>st</sup> to 2 <sup>nd</sup> visit (days)	6.8 $\pm$ 0.4	5.1 $\pm$ 4.8	7.1 $\pm$ 1.5	7.1 $\pm$ 1.0	7.0 $\pm$ 4.2
Interval: 1 <sup>st</sup> to 3 <sup>rd</sup> visit (days)	13.3 $\pm$ 1.4	10.2 $\pm$ 7.0	16.6 $\pm$ 7.9	14.1 $\pm$ 0.8	104.4 $\pm$ 9.6
Standard VF MD (dB): mean $\pm$ SD (range)	2.2 $\pm$ 1.1 (–0.2 to 3.9)	0.2 $\pm$ 1.2 (–2.7 to 3.2)	0.5 $\pm$ 1.6 (–2.8 to 3.3)	–3.6 $\pm$ 4.1 (–12.0 to 1.9)	–3.7 $\pm$ 2.3 (–9.4 to 1.2)
SWAP MD (dB): mean $\pm$ SD (range)	0.6 $\pm$ 1.7 (–2.4 to 4.1)	–1.0 $\pm$ 2.8 (–8.5 to 3.5)	–0.4 $\pm$ 3.2 (–9.9 to 5.6)	–4.3 $\pm$ 4.3 (–13.8 to 2.9)	–4.7 $\pm$ 4.5 (–16.3 to 3.3)

SWAP VFs were performed on the same visit day, and the order in which they were performed was randomized between patients and visits. A short break separated the two VF exams. In addition, rest periods were provided approximately halfway through a given test, as well as at the patient's request or technician's discretion.

The relatively short interval, as well as overall clinical stability of the disease, as assessed by a glaucoma expert, made true glaucomatous progression of meaningful magnitude in these groups unlikely. The normal II group performed 3 standard and 3 SWAP VFs within 2–8 weeks. The patients in the glaucoma II group performed the same series of VFs over a period of 3 to 4 months (Table 1). Exclusion of true glaucomatous progression in this group was addressed in a quantitative approach as follows: in addition to close clinical monitoring, an in-depth analysis of the 1-, 2-, 3-, 6-, and 12-month standard and SWAP VFs was undertaken.<sup>3,4,8</sup> Briefly, the 6- and 12-month VFs were analyzed based on the criteria listed below, to insure that meaningful repeatable glaucomatous progression did not take place. These criteria were used separately for standard VF and for SWAP.

Progression of visual field defects for the glaucoma II group was defined as either (1) development of a new scotoma, defined as 2 adjacent points in a previously normal area, at the 0.01 probability level on the pattern deviation plot, or one point within the central 10° that declined by  $\geq 10$  dB; (2) expansion of existing scotoma, defined as 2 contiguous points adjacent to an existing scotoma that declined by  $\geq 10$  dB; (3) deepening of an existing scotoma, defined as 2 points in an existing scotoma that declined by  $\geq 10$  dB. A baseline was created based on a point-by-point threshold average of the first 2 visual fields taken 1 week apart. Those patients with repeatable progression on 2 consecutive VFs were excluded. Only the VFs obtained during the first 3 to 4 months were used in the TRV calculations for this group. These progression criteria were developed by a group of glaucoma specialists for a longitudinal study where progression of visual fields was a primary outcome measure. Visual field data for the normal I, suspect and glaucoma I groups were collected at University of California, San Diego, for the normal II group at Yale University (by JC) and data for the glaucoma II group were acquired during the course of a multi-center study in which patients received commonly used pressure-lowering medications. LTF results were previously published for group II normals.<sup>14</sup> Reliability criteria included fixation losses  $\leq 25\%$ , false positive responses  $\leq 25\%$  and false negative responses  $\leq 25\%$ . Whenever a field did not meet the above reliability criteria, the field was repeated within 2 weeks and was included only if it then met the reliability criteria.

## Calculating TRV

Visual field absolute threshold values were exported from the HFA using Peridata v7.2 (Peridata Software GmbH; Huerth, Germany).

Test–retest variability was calculated as the standard deviation of the 3 threshold dB values for a given location, taken from 3 VFs. When a specific location was retested during a VF testing session, the average of the 2 thresholds was used in the analysis. Only this average value is the value exported by Peridata. To ascertain whether using the average would hide any differences found in points that are double-determined, the 10 pre-defined locations of the 24-2 grid that are always tested twice were incorporated into the model.

The global long-term fluctuation is often defined as the difference between the total variability in threshold sensitivity over time and the variability due to short-term fluctuation.<sup>18–20</sup> In this study, we chose to calculate the location-specific TRV, the total fluctuation over time (which contains the short-term fluctuation), because TRV is a direct measure of change over time. Using HFA program 24-2, short-term fluctuation is routinely determined at only 10 predetermined locations within the entire visual field, and even in these locations its estimation is based on only 2 repetitions.<sup>21</sup> Hence, a reliable point-by-point short-term fluctuation value is not routinely acquired. Attempting to substitute the global field short-term fluctuation value as the short-term fluctuation value for each individual point would not be reasonable, since short-term fluctuation is location dependent.<sup>22</sup> While not addressed in this study, short-term fluctuation was previously estimated at 25% of the total TRV.<sup>10</sup>

## VF Analysis

For each individual, 3 standard VF and 3 SWAP 24-2 full-threshold fields were used to calculate the TRV at each visual field test location. Severity of glaucomatous VF damage (age-corrected threshold-depression) was defined as the total deviation value at each test location from the age-corrected expected value. The conversion was performed separately for standard VF and SWAP. The expected values were derived from a multi-center normative database containing 348 normal subjects, tested on both standard VF and SWAP. This normative database was chosen because it contains identical subjects for standard VF and SWAP, as opposed to the separate databases (standard VF/SWAP) provided with the Humphrey VF machines.

Locations with a “<0 dB” value were excluded: the maximum brightness a projected target can attain in the Humphrey Field Analyzer is defined as 0 dB. Severely depressed retinal locations at which the 0 dB stimulus is not perceived are given a threshold value of “<0”. In fact, “<0” thresholds span a wide range of retinal sensitivities, from –1 dB to no light perception. Therefore, any attempt to uniformly substitute <0 threshold with any single numerical value would significantly underestimate the true magnitude of the TRV at that location by an unknown amount. For this reason, all locations that scored “<0” even once in any of the 3

Table 2. Quantitative Characteristics of LTV (in dB  $\pm$  SD) for Normals, Glaucoma Suspects and Glaucoma Patients

	Normals I (n = 10)	Normals II (n = 31)	Suspects (n = 10)	Glaucoma I (n = 10)	Glaucoma II (n = 25)
For standard VF					
Overall LTV	1.26 $\pm$ 0.94	1.28 $\pm$ 0.85	1.53 $\pm$ 1.04	2.34 $\pm$ 2.38	2.33 $\pm$ 1.87
Superior hemifield	1.40 $\pm$ 1.00	1.35 $\pm$ 0.90	1.58 $\pm$ 1.00	2.16 $\pm$ 2.07	2.56 $\pm$ 2.02
Inferior hemifield	1.12 $\pm$ 0.84	1.22 $\pm$ 0.79	1.49 $\pm$ 1.08	2.50 $\pm$ 2.61	2.10 $\pm$ 1.69
Nasal hemifield	1.22 $\pm$ 0.88	1.27 $\pm$ 0.87	1.45 $\pm$ 0.94	2.37 $\pm$ 2.34	2.38 $\pm$ 1.90
Temporal hemifield	1.31 $\pm$ 0.99	1.30 $\pm$ 0.82	1.63 $\pm$ 1.42	2.30 $\pm$ 2.42	2.27 $\pm$ 1.84
1 tested threshold	1.28 $\pm$ 0.97	1.31 $\pm$ 0.86	1.56 $\pm$ 1.02	2.39 $\pm$ 2.36	2.43 $\pm$ 1.93
2 tested thresholds averaged	1.18 $\pm$ 0.77	1.18 $\pm$ 0.79	1.40 $\pm$ 1.12	2.14 $\pm$ 2.44	1.89 $\pm$ 1.53
For SWAP					
Overall LTV	2.00 $\pm$ 1.39	1.83 $\pm$ 1.34	2.45 $\pm$ 2.13	2.05 $\pm$ 1.32	2.90 $\pm$ 1.56
Superior hemifield	2.36 $\pm$ 1.56	2.00 $\pm$ 1.45	2.63 $\pm$ 2.12	2.14 $\pm$ 1.40	3.19 $\pm$ 1.94
Inferior hemifield	1.66 $\pm$ 1.09	1.65 $\pm$ 1.20	2.28 $\pm$ 2.14	1.97 $\pm$ 1.24	2.63 $\pm$ 1.93
Nasal hemifield	2.02 $\pm$ 1.46	1.89 $\pm$ 1.42	2.56 $\pm$ 2.23	2.11 $\pm$ 1.31	2.93 $\pm$ 1.94
Temporal hemifield	1.99 $\pm$ 1.31	1.76 $\pm$ 1.23	2.33 $\pm$ 2.01	1.98 $\pm$ 1.32	2.86 $\pm$ 1.97
1 tested threshold	2.10 $\pm$ 1.43	1.87 $\pm$ 1.36	2.52 $\pm$ 2.16	2.06 $\pm$ 1.33	3.00 $\pm$ 1.98
2 tested thresholds averaged	1.58 $\pm$ 1.14	1.66 $\pm$ 1.26	2.17 $\pm$ 2.01	2.01 $\pm$ 1.26	2.49 $\pm$ 1.83

VFs, for a given eye, were eliminated. On average, for standard VF, the number of locations eliminated was 0.02 for normals, 0 for suspects and 1.86 for glaucoma. For SWAP the number of locations eliminated for normal, suspect, and glaucomatous eyes was 0.24, 0.5 and 4.29 respectively.

## Statistical Analysis

A multiple regression model was run separately for standard VF and for SWAP, incorporating the following factors:

1. Severity of VF damage, expressed as total deviation, defined as the difference (in dB) of the mean threshold value from the expected threshold value (at each VF location) in an age-matched normal population.
2. Eccentricity, measured in degrees from fixation.
3. Diagnosis (normal [n = 41], glaucoma suspect [n = 10], glaucoma [n = 35]).
4. Study group, summarizing differences between groups having a similar diagnosis. This factor summarizes differences between the 2 normal and the 2 glaucoma subgroups, which are not accounted for by differences in diagnosis. Such differences include varying intervals between test sessions in the different study groups (see Table 1) as well as possible effects related to different testing techniques or patient selection.
5. "Superior vs. inferior" hemifield.
6. "Nasal vs. temporal" hemifield.
7. "Location (overall)," the actual coordinates of each point in the 24-2 grid. Location (overall) is hence inclusive of eccentricity, "superior vs. inferior," "nasal vs. temporal," and once-tested against twice-tested locations.
8. "One versus two thresholds," defined as locations tested once versus twice. Since 10 predefined locations are always tested twice, and averaged, this factor accounts for any reduction in TRV stemming from double determination of the threshold.
9. Age.
10. "Pupil size variability," defined as the standard deviation (SD) of the three measurements.
11. "Pupil size mean," defined as the mean of the three pupil measurements.
12. "Between subject variation" in the overall level of TRV of each subject, i.e., each subject is allowed a distinct inter-

cept (modeled as random effect) that raises or depresses all values.

13. "Residual variation," i.e., the remaining variability not attributable to any of the above 12 factors.

This linear, mixed effects model<sup>23</sup> was fitted to the data using the Splus (MathSoft, Inc., Cambridge, MA) function 'lme'. Preliminary inspection of the data using normal probability (quantile-quantile) plots suggested that the TRV residuals have a long right tail. This was approximately corrected by taking the square root of TRV. In the analyses that follow, both TRV and its square root were studied. The results were qualitatively similar, and for ease of interpretation only results based on TRV are presented. All *P* values are based on the likelihood ratio statistic; tabulated estimates give the maximum likelihood solutions.

An automated pupil-measuring device (enabling measurements in 0.1 mm increments) exists only on the newer HFA II units. Accurate within-the-bowl pupil measurements, in 0.5 mm increments, were available for 10 normal, 10 suspect and 10 glaucoma subjects (groups Normal I, Suspect I and Glaucoma I), obtained as follows: a calibrated template card was prepared so that the actual pupil size, as seen on the HFA I screen, could be assessed against this template, while in the bowl, at the standardized background illumination. Two summaries of pupil size were incorporated into the statistical model: first, variations in pupil size between the 3 tests, calculated as the SD. Second, mean pupil size was used in the model independent from pupil size SD, to ascertain whether individuals with larger (or smaller) pupils (even if no size change was observed between tests) have higher TRV. Pupil size data were run separately on the subset of 30 subjects for whom accurate pupil data were available.

## Results

Table 1 presents demographic data by groups. Although all patients enrolled in this study had prior visual field experience, they were tested for a residual learning curve effect separately for standard VF and SWAP. For each of the 5 groups, the MD of the first, second and third VFs was compared using ANOVA. No statistically significant differences were found between the 3 sets of VFs in any of the subgroup analyses (ANOVA *P* values range, 0.29–0.97). Global TRV data for each of the 5 subgroups are presented in Table 2.

Table 3. Sources of Variability Quantified

I. Standard VF				
	Variance Explained (dB <sup>2</sup> )	Proportion of Variance Explained (Percentage) <sup>c</sup>	P Value <sup>d</sup>	For Each Factor: Which Item Was Associated with Higher LTV?
Age	0.002	0.1	0.73	younger>older
Study group <sup>a</sup> :	0.069	3.5	<b>0.0054</b>	GII>NI>S>GI>NII
Diagnosis (N/S/G)	0.056	2.9	0.013	S>G>N
Total deviation	0.30	15.5	<b>&lt;0.0001</b>	defective>normal
Location (overall) <sup>b</sup> :	0.053	2.7	<b>&lt;0.0001</b>	
Nasal vs. temporal	0.0012	0.06	0.087	temporal>nasal
Superior vs. inferior	0.0033	0.17	<b>&lt;0.0001</b>	superior>inferior
Location tested once vs. twice	0.0008	0.04	0.043	tested once>twice
Eccentricity	0.021	1.1	<b>&lt;0.0001</b>	peripheral>central
Between-subject variation	0.16	8.0	<b>&lt;0.0001</b>	
Residual variation	1.19	61.0	<b>&lt;0.0001</b>	
Mean standard LTV variance for all subjects = 1.94 dB <sup>2</sup> .				
II. SWAP				
	Variance Explained (dB <sup>2</sup> )	Proportion of Variance Explained (Percentage) <sup>c</sup>	P Value <sup>d</sup>	For Each Factor: Which Item Was Associated with Higher LTV?
Age	0.0016	0.06	0.77	older>younger
Study group <sup>a</sup> :	0.18	7.0	<b>&lt;0.0001</b>	GII>NI>GI>S>NII
Diagnosis (N/S/G)	0.15	5.9	<b>0.0037</b>	G>S>N
Total deviation	0.18	6.9	<b>&lt;0.0001</b>	defective>normal
Location (overall) <sup>b</sup> :	0.1067	4.1	<b>&lt;0.0001</b>	
Nasal vs. temporal	0.0004	0.02	0.074	nasal>temporal
Superior vs. inferior	0.045	1.7	<b>&lt;0.0001</b>	superior>inferior
Location tested once vs. twice	0.0042	0.16	<b>0.0009</b>	tested once>twice
Eccentricity	0.017	0.64	<b>&lt;0.0001</b>	peripheral>central
Between-subject variation	0.35	13.5	<b>&lt;0.0001</b>	
Residual variation	1.73	66.6		
Mean SWAP LTV variance for all subjects = 2.59 dB <sup>2</sup> .				
<sup>a</sup> Study group subsumes Diagnosis.				
<sup>b</sup> Location (overall) subsumes Nasal vs. Temporal, Superior vs. Inferior, Location tested once vs. twice and Eccentricity.				
<sup>c</sup> Totals of variance explained (90.8% for standard VF and 98.2% for SWAP) do not equal 100% due to correlation between factors (especially Total deviation and Study group for standard VF).				
<sup>d</sup> P values of <0.01 are presented in bold text.				

Table 3 presents the results of a model including all 86 study subjects. Data are presented separately for standard VF and for SWAP. Of the factors included in the table (excluding between-subject and residual variation), total deviation (severity) exerted the largest effect on TRV, particularly for standard VF. Ranked next were diagnosis and location for SWAP. All other factors, although mostly statistically significant (see Table 3, rightmost column), did not contribute much to the variability in TRV in the population studied. In fact, all factors combined (excluding between-subject and residual variation) accounted for less than a

third of the overall variability seen in TRV in both the standard VF and the SWAP models.

Pupil size for subjects in this study was 4.0 ± 0.85 mm (range, 3.0 to 6.0; subjects with pupil size <3.0 mm were excluded from this study). Variability in pupil size (SD of the 3 measurements ± SD) was 0.27 ± 0.26 (range, 0 to 0.58). A separate statistical model was run on a subset of 30 subjects who had reliable pupil data. This model incorporated “pupil size variability” and “pupil size mean” as additional factors (Table 4). In this model one pupil size was not found to be statistically significant, at the P < 0.01

Table 4. A Separate Model That Included the Effect of Pupil Size on LTV Along with All Other Factors

VF type	Factor	Variance Explained (dB <sup>2</sup> )	Proportion of Variance Explained (Percentage)	P Value
Standard VF	Pupil size- variability	0.043	3.2	0.041
Standard VF	Pupil size- mean	0.0080	0.59	0.26
SWAP	Pupil size- variability	0.048	2.8	0.042
SWAP	Pupil size- mean	0.0018	0.10	0.59

Normals, n = 10; suspects, n = 10; glaucoma, n = 10.

level. However, with respect to the magnitude of the effect, a modest effect was found for "pupil size variability" on TRV, in the range of 3% for both standard VF and for SWAP. After accounting for variations in pupil size from one exam to the next, the "pupil size mean" contributed little to variability of TRV. This, in turn, implies that when pupil size is kept stable across tests, it seems to make little difference whether the pupil is uniformly large or uniformly small (when in the range of 3–6 mm).

## Discussion

This study was undertaken to analyze which factors contribute to the variability in visual fields, and to what extent. While many of these factors are known to affect variability in a statistically significant manner,<sup>3–6,8,10,13,15</sup> little information is available about the relative magnitude of their individual effect. A quantitative analysis may help establish which of these factors needs to be accounted for in progression algorithms, and in contrast, what proportion the remaining true physiological variation component encompasses.

In order to better isolate TRV, two issues had to be addressed: minimizing the learning effect,<sup>24–26</sup> and exclusion of true progression. All subjects performed at least 2 pre-study fields to minimize the learning curve effect. In addition, a formal statistical analysis excluded a consistent trend in the global mean defect value for the 3 serial VFs. Exclusion of true progression was addressed in two ways. Suspects and Glaucoma I patients were tested during a very short interval, while patients showing progression were excluded from the glaucoma II group by way of an in-depth analysis of the 1, 2, 3, 6 and 12 month VFs, separately for standard VF and SWAP.<sup>15</sup>

Most of the factors studied were found to exert a statistically significant effect on TRV. This is in agreement with the published literature on variability in automated static perimetry.<sup>3–5,7,8,10,22</sup>

A multiple regression model determined that none of these factors contributed extensively to the variability encountered in the visual field thresholds over time, for either standard VF or for SWAP. It thus appears that while statistical significance is indeed present (as shown again in this study) for most of the factors, the magnitude of the effect is small.

Table 3 provides an interesting insight into the relative contribution of the studied factors to explaining the variability. Diagnosis, a factor known to significantly underlie variability,<sup>3,4,6,19</sup> surprisingly explained only 2.9% and 5.9% of the variation of the model, respectively, for standard VF and SWAP. An even more striking example is the case of eccentricity, which was previously shown to be significantly related to variability in several studies,<sup>3–5,8</sup> as well as in this study, with a  $P$  value of  $<0.0001$  for both standard VF and SWAP. However, the amount of the variation attributable to eccentricity in the TRV model was a mere 1.1% and 0.64% respectively, for standard VF and SWAP.

In contrast, defect severity at any given location was the single largest factor associated with increased TRV. Severity was found to have a much larger effect on TRV than the

diagnosis status of the individual tested. Hence, those VF locations still showing normal values in glaucoma patients also showed TRV values that were comparable to values found in normal subjects. This is a reassuring finding in a data set comprised of multiple subject groups from several sites, and with different inter-session intervals (see Table 1).

We wondered if the number of suspects was large enough to support the finding that the diagnostic group played a limited role in explaining TRV. There are two aspects to this issue: 1) whether the estimated coefficients are accurate enough to rule out substantial differences among diagnostic groups and 2) whether the observed differences would amount to substantial variance explained in a population with more suspects. To address this issue, a 95% confidence interval for the percent of variance explained due to diagnostic group (using non-central F distribution and the observed values of the conditional F-statistics) was first constructed. The upper endpoint of this interval for standard visual fields was 10.6% and that for SWAP was 13.6%. Thus, diagnostic group explains at most a minor share of the variation in TRV in populations with similar proportions of suspects, normals and glaucoma patients. Second, the amount of variability observed in suspects was intermediate between normals and glaucoma patients for SWAP. For standard visual fields, the observed variation in suspects was similar to that in glaucoma patients (and nonsignificant at  $P = 0.97$ ) which in turn was substantially greater than that seen in normals, so increasing the proportion of suspects would either slightly increase or decrease the variance depending on whether the proportion of glaucoma patients or normals was decreased by a like amount. Likewise, Flammer found that all components of fluctuation for suspects fell in between normals and glaucoma patients.<sup>20</sup>

Location (overall), although significantly associated with TRV ( $P < 0.0001$ ), surprisingly contributed very little to the model, with 2.7% and 4.1%, respectively, for standard VF and SWAP. Not only did none of the global location parameters (eccentricity, "superior vs. inferior," "nasal vs. temporal," and locations tested once vs. twice) contribute much, but even after accounting for each location in the 24-2 VF grid, only a small proportion of the overall TRV variability could be accounted for. In our data set, age and "nasal vs. temporal" did not show a statistically significant effect on variability in TRV. Similarly, the magnitude of effect for these two factors is negligible. Previous studies reported the significance of the association between location/eccentricity and TRV, but did not assess the magnitude of the effect. An important advantage of the present study is the determination of the relative magnitude of each effect on TRV.

Between-subject variation accounted for 8.0% and 13.5% for standard VF and SWAP, respectively. These moderately sized values hint that while differences unique to each individual exist, they are not overwhelming in magnitude. This is somewhat reassuring for future refinement of progression algorithms.

In contrast, "residual variation" was by far the most disturbing component of our model. At 61.0% and 66.6% for standard VF and SWAP, respectively, this component

seriously undermines our optimism in the ability to reduce the variability of various field tests and to fine-tune progression algorithms. Future research should be directed towards identifying additional factors that are predictive of TRV. Probable major candidates are the variability inherent in psychophysical procedures, the intrinsic physiological variability present in both healthy and diseased eyes, and patient reliability. We should mention, however, that some of the residual unexplained variation may be due to use of only 3 visual fields in the series. We chose 3 values for calculation of TRV as a compromise between more data on the one hand, and a practical approach that stands a chance in the clinical arena, on the other hand. Since our primary intention was to analyze TRV in a way that would later help devise algorithms of progression on an individual basis, we chose 3 values as the minimal number of repetitions that are reasonable from a statistical point of view, while still manageable in the clinical setting.

Concerning differences in the magnitude of TRV between standard VF and SWAP, a previous study based on our glaucoma II group ( $n = 25$ ) showed SWAP TRV to be 0.55 dB higher than standard VF TRV for glaucoma patients.<sup>15</sup> In the current data ( $n = 86$ ) SWAP TRV was found to be higher than standard VF TRV by 0.59 ( $P < 0.0001$ ), 0.33 ( $P < 0.0001$ ) and 0.49 ( $P < 0.0001$ ) dB for normal subjects, glaucoma suspects and glaucoma patients, respectively. It appears that this difference in TRV between standard VF and SWAP remains, regardless of diagnostic category.

No large differences in the relative contribution of the factors underlying variability in TRV were noted between the standard VF and the SWAP models. Factors contributing relatively more weight in the standard VF model included "total deviation" and "eccentricity," while "diagnosis," "location" (especially "superior-inferior"), and "between subject variation" contributed somewhat more in the SWAP model.

As an integral part of the HFA full-threshold testing algorithm, 10 pre-defined locations are always tested twice. This information is the basis for the short-term fluctuation calculation presented on the HFA printout. Exported data, as well as further HFA statistical analysis, both use the average threshold value for those twice-tested locations. We chose, in this study, to use the average value for any double-determination points, even though TRV might be somewhat underestimated. The hypothesis that this averaging may tend to remove a portion of the short-term fluctuation, and hence serve to reduce the computed TRV, was tested. Surprisingly, only 0.04% of the total variability for standard VF and 0.16% for SWAP could be accounted for by the fact that certain locations are uniformly tested twice. It is worthwhile to note, however, that these results do not pertain to those locations tested twice owing to a "suspicious" threshold as determined by the automated algorithm.

Two aspects of pupil size can potentially affect variability: variability of pupil size from one test to the next, and a large pupil, even if constant from one test to the next. Our results show the former to be far more important, but that neither exerts a meaningful effect on the overall TRV. This, in turn, substantiates the accepted concept that fluctuations

in pupil size within a reasonable range (3 to 6 mm) do not need to be accounted for in progression algorithms since little effect on the overall contrast is expected.<sup>1</sup>

A recent study<sup>6</sup> analyzed the variability seen in frequency of seeing data collected from subjects with glaucoma, ocular hypertension, optic neuritis and normals. Regardless of the cause of ganglion cell loss, they were able to show a similar relationship between response variability and sensitivity. They concluded that response variability might be dependent on functional ganglion cell density regardless of the etiology of the cell loss. For these frequency of seeing curves, generated by over 120 presentations per location, severity was shown to account for much of the response variability encountered in the data ( $R^2 = 0.57$ ). Interestingly, although the magnitude of the effect was small, our study confirms that the single most influential factor on TRV, in any given individual, is indeed sensitivity loss (total deviation).

Three values for calculation of TRV were chosen as a reasonable compromise between more data (leading to a more accurate determination of TRV with tighter confidence intervals) on the one hand, and a practical approach that stands a chance in the clinical arena, on the other. Our primary intention was to analyze TRV in a way that would later help devise algorithms of progression. If anything, this choice of 3 repetitions will make identification of progression more conservative. On the other hand, we are aware that calculating TRV based on only 3 repetitions might have resulted in less robust estimation of TRV. It is possible that a larger number of VF repetitions used for the calculation of TRV might reduce the magnitude of the unexplained variability found in this study.

In summary, while most of the factors in the model exert a statistically significant measurable effect on TRV, they were found to contribute modestly to the overall variability noted. In fact, all 11 factors combined accounted for less than 31% of the TRV variability seen in the standard VF model and less than 20% seen in the SWAP model. Over 70% of the variability remained unaccounted for. This explains the difficulty in finding methods to identify true confirmed change in visual fields due to glaucomatous progression. Some variability will always be present when testing human vision. However, some proportion of this remaining variability needs to be eliminated by improving testing techniques and determining if there are other contributing factors under our control. In the meantime, any change noted in visual fields should be confirmed on more than one occasion to increase the likelihood that the change is truly due to progressing glaucoma.

## References

1. Anderson DR, Patella VM. Automated Static perimetry, 2nd ed. St. Louis: Mosby, 1999:45-79.
2. Hutchings N, Wild JM, Hussey MK, et al. The long-term fluctuation of the visual field in stable glaucoma. *Invest Ophthalmol Vis Sci* 2000;41:3429-36.
3. Heijl A, Lindgren A, Lindgren G. Test-retest variability in glaucomatous visual fields. *Am J Ophthalmol* 1989;108:130-5.

4. Boeglin RJ, Caprioli J, Zulauf M. Long-term fluctuation of the visual field in glaucoma. *Am J Ophthalmol* 1992;113:396–400.
5. Chauhan BC, Johnson CA. Test-retest variability of frequency-doubling perimetry and conventional perimetry in glaucoma patients and normal subjects. *Invest Ophthalmol Vis Sci* 1999;40:648–56.
6. Henson DB, Chaudry S, Artes PH, et al. Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Invest Ophthalmol Vis Sci* 2000;41:417–21.
7. Flammer J, Drance SM, Augustiny L, Funkhouser A. Quantification of glaucomatous visual field defects with automated perimetry. *Invest Ophthalmol Vis Sci* 1985;26:176–81.
8. Wild JM, Moss ID, Whitaker D, O'Neill EC. The statistical interpretation of blue-on-yellow visual field loss. *Invest Ophthalmol Vis Sci* 1995;36:1398–410.
9. Haas A, Flammer J, Schneider U. Influence of age on the visual fields of normal subjects. *Am J Ophthalmol* 1986;101:199–203.
10. Katz J, Sommer A. A longitudinal study of the age-adjusted variability of automated visual fields. *Arch Ophthalmol* 1987;105:1083–6.
11. Lindenmuth KA, Skuta GL, Rabbani R, Musch DC. Effects of pupillary constriction on automated perimetry in normal eyes. *Ophthalmology* 1989;96:1298–301.
12. Lindenmuth KA, Skuta GL, Rabbani R, et al. Effects of pupillary dilation on automated perimetry in normal patients. *Ophthalmology* 1990;97:367–70.
13. Rebolleda G, Munoz FJ, Fernandez Victorio JM, et al. Effects of pupillary dilation on automated perimetry in glaucoma patients receiving pilocarpine. *Ophthalmology* 1992;99:418–23.
14. Kwon YH, Park HJ, Jap A, et al. Test-retest variability of blue-on-yellow perimetry is greater than white-on-white perimetry in normal subjects. *Am J Ophthalmol* 1998;126:29–36.
15. Blumenthal EZ, Sample PA, Zangwill L, et al. Comparison of long-term variability for standard and short-wavelength automated perimetry in stable glaucoma patients. *Am J Ophthalmol* 2000;129:309–13.
16. Katz J, Gilbert D, Quigley HA, Sommer A. Estimating progression of visual field loss in glaucoma. *Ophthalmology* 1997;104:1017–25.
17. Chylack LT Jr, Wolfe JK, Singer DM, et al. The lens opacities classification system III. The longitudinal study of cataract study group. *Arch Ophthalmol* 1993;111:831–6.
18. Bebie H, Fankhauser F, Spahr J. Static perimetry: accuracy and fluctuations. *Acta Ophthalmol (Copenh)* 1976;54:339–48.
19. Flammer J, Drance SM, Zulauf M. Differential light threshold. Short- and long-term fluctuation in patients with glaucoma, normal controls, and patients with suspected glaucoma. *Arch Ophthalmol* 1984;102:704–6.
20. Flammer J, Drance SM, Schulzer M. Covariates of the long-term fluctuation of the differential light threshold. *Arch Ophthalmol* 1984;102:880–2.
21. Chauhan BC, LeBlanc RP, Drance SM, et al. Effect of the number of threshold determinations on short-term fluctuation in automated perimetry. *Ophthalmology* 1991;98:1420–4.
22. Heijl A, Lindgren G, Olsson J. Normal variability of static perimetric threshold values across the central visual field. *Arch Ophthalmol* 1987;105:1544–9.
23. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982;38:963–74.
24. Heijl A, Lindgren G, Olsson J. The effect of perimetric experience in normal subjects. *Arch Ophthalmol* 1989;107:81–6.
25. Heijl A, Bengtsson B. The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol* 1996;114:19–22.
26. Marra G, Flammer J. The learning and fatigue effect in automated perimetry. *Graefes Arch Clin Exp Ophthalmol* 1991;229:501–4.