Glaucoma

Long- and Short-Term Variability of Perimetry in Glaucoma

Stuart K. Gardiner[1], William H. Swanson[2], and Steven L. Mansberger[1]

[1] Legacy Devers Eye Institute, Legacy Health, Portland, OR, USA
[2] School of Optometry, Indiana University, Bloomington, IN, USA

Purpose: Test–retest variability in perimetry consists of short-term and long-term components, both of which impede assessment of progression. By minimizing and quantifying the algorithm-dependent short-term variability, we can quantify the algorithm-independent long-term variability that reflects true fluctuations in sensitivity between visits. We do this at locations with sensitivity both < 28 dB (when the stimulus is smaller than Ricco's area and complete spatial summation can be assumed) and > 28 dB (when partial summation occurs).

Methods: Frequency-of-seeing curves were measured at four locations of 35 participants with glaucoma. The standard deviation of cumulative Gaussian fits to those curves was modeled for a given sensitivity and used to simulate the expected short-term variability of a 30-presentation algorithm. A separate group of 137 participants was tested twice with that algorithm, 6 months apart. Long-term variance at different sensitivities was calculated as the LOESS fit of observed test–retest variance minus the LOESS fit of simulated short-term variance.

Results: Below 28 dB, short-term variability increased approximately linearly with increasing loss. Long-term variability also increased with damage below this point, attaining a maximum standard deviation of 2.4 dB at sensitivity 21 dB, before decreasing due to the floor effect of the algorithm. Above 30 dB, the observed test–retest variance was slightly smaller than the simulated short-term variance.

Conclusions: Long-term and short-term variability both increase with damage for perimetric stimuli smaller than Ricco's area. Above 28 dB, long-term variability constitutes a negligible proportion of test–retest variability.

Translational Relevance: Fluctuations in true sensitivity increase in glaucoma, even after accounting for increased short-term variability. This long-term variability cannot be reduced by altering testing algorithms alone.

Introduction

In clinical care for patients with glaucoma, accurately measuring the rate at which the patient's functional loss is progressing is vital for making appropriate treatment decisions.[1] Many eyes progress slowly, whereas a few progress rapidly enough that the patient is at risk of visual impairment or blindness within their expected lifespan.[2] However, the crucial task of measuring this rate is hampered by the substantial test–retest variability of standard automated perimetry.[3,4] Thus, accurate rate measurements require either very frequent testing[5] (which is inconvenient for patients and resisted by payers) or long follow-up durations[6] (during which more disease progression toward blindness may have occurred). In order to reduce this variability and hence aid assessment of progression, it is essential to actually understand the sources of the variability. This understanding can guide efforts to optimize testing and also uncover upper bounds on the theoretically achievable repeatability for a perfect observer.[7] There are two main types of variability that affect visual field measurements: short term and long term.[8,9] Short-term variability is generally understood to represent the test–retest variability within a single session, which perimetric techniques and algorithms can seek to minimize. We hypothesize that there is

also substantial long-term variability, which would remain even with a perfectly repeatable test and that its magnitude may be affected by disease severity.

We define short-term variability as the variance of sensitivity estimates that would be expected from repeated testing on a single test session, using two identical, independent, interleaved testing algorithms (i.e., equalizing factors such as fatigue and alertness so the difference between the two sensitivity estimates can be attributed to natural response variability). Clearly this short-term variability is dependent on the testing algorithm used. Thus, we first measured frequency-of-seeing curves (i.e., psychometric function) using the method of constant stimuli.[10,11] The detection threshold in clinical perimetry is defined as the contrast at which the subject responds to 50% of stimulus presentations, and contrast sensitivity is the reciprocal of this contrast.[12] However, stimulus detection is dependent on the number of discrete neural spikes produced by retinal ganglion cells within a short window of time. The exact timing of spikes, and so the number within that window, varies even when identical visual stimuli are presented.[13–17] Hence, the frequency-of-seeing curve is not a step function but a gradual sigmoidal increase in response probability with contrast, which can be fit by, for example, a cumulative Gaussian distribution. We can then derive not only an accurate measurement of contrast sensitivity but also the standard deviation (SD) of this fitted cumulative Gaussian distribution, a measure of response variability.[18] This allows us to predict the short-term variability for any chosen testing algorithm, for a given "true" underlying sensitivity, by repeated simulated runs of the algorithm. Notably, this SD (and hence the short-term variability) increases markedly at locations with lower sensitivity.[10]

We define long-term variability as the variance in the true sensitivity between days (i.e., as if it were measured using a perfect testing algorithm with no measurement variability and zero short-term variability).[9,19] Some previous studies have defined long-term variability simply as the test–retest variability with a gap of weeks or months between the tests.[8] By our definitions, the total test–retest variance from patient data equals the sum of the short-term and long-term variances, which are independent of one another. There are various sources of long-term variability, such as time of year,[20] time of day,[21] technician experience,[21] and level of fatigue at the time of testing.[22,23] It has been shown that variability between fields taken annually is greater than variability between fields taken weekly,[24] suggesting the presence of longer term fluctuations in true function rather than just the effect of test reliability on a given day. That study also suggested

an increase in long-term variability with glaucomatous damage, raising the possibility of underlying variations in true disease status (not just testing variability).

Recently, research has revealed that certain characteristics of sensitivity estimates from standard automated perimetry, including variability, fundamentally change when sensitivity falls below a certain level. In most healthy eyes, the size III stimulus that is used in most clinical care is larger than Ricco's area of complete spatial summation.[25] This means that partial spatial summation occurs; if stimulus area is doubled, the detection threshold (the reciprocal of contrast sensitivity) will decrease but by less than half. In glaucoma, as sensitivity decreases, Ricco's area enlarges.[26] When sensitivity is below around 28 dB, the size III stimulus has been found to be smaller than Ricco's area at most locations, and complete spatial summation occurs. As long as the stimulus area stays within Ricco's area, doubling the stimulus area will halve the detection threshold. It has been suggested that using stimuli smaller than Ricco's area will increase the signal-to-noise ratio of perimetry.[27] It has also been suggested that, when stimuli are larger than Ricco's area, variability increases much less than when stimuli are smaller than Ricco's area.[28] Therefore, it becomes important to quantify and to better understand the causes of variability both above and below this level of damage, with the realization that the relation between sensitivity and variability may not be homogeneous.

In order to quantify the underlying long-term variability, we need to first quantify the short-term variability. Then, the long-term variance (i.e., the square of the SD) over an extended time interval equals the test–retest variance over that interval minus the short-term variance. In Experiment 1, by reanalysis of data from a previous experiment,[29,30] we determined the average standard deviation of a frequency-of-seeing curve for a given contrast sensitivity in patients with glaucoma. This builds on the work of Henson et al.[10] to include more severely damaged locations; extrapolating their model to <10 dB, beyond their measurements, produces unrealistically high predictions of variability.[31] We also extended the results to include testing with a larger size V (1.72° diameter) stimulus,[30] which has been reported to reduce variability.[32] In Experiment 2, we determined the test–retest variability of a custom-written, high-accuracy testing algorithm on a separate group of patients with suspected or confirmed glaucoma. This algorithm minimizes short-term variability by using 30 stimulus presentations per location, compared with around five per location in the Swedish Interactive Thresholding Algorithm (SITA) Standard algorithm,[33] and

by using size V stimuli.[32] In Experiment 3, the short-term variability for our high-accuracy algorithm was predicted from simulations, using results from Experiment 1. By both minimizing the short-term variability and predicting the remaining short-term variability, we aimed to accurately characterize the long-term variability at different severities of glaucomatous damage. This can lead to improved understanding of the causes of that variability, in addition to informing the development of improved diagnostic testing.

Methods

Experiment 1. Characterizing Short-Term Variability

The first part of the study aimed to quantify and characterize short-term variability at different levels of glaucomatous loss by measuring frequency-of-seeing curves. For this, data were taken from a previously published study, and full details of the experiment have been reported.[29,30] In brief, 35 participants with moderate to severe primary open-angle glaucoma, as determined by their clinician, were recruited from the Devers Eye Institute glaucoma clinic. For eligibility, participants were required to have two or more non-adjacent visual field locations with sensitivities between 6 and 18 dB on both of their two most recent clinic visits (Humphrey Field Analyzer with 24-2 test pattern, size III stimulus, and SITA Standard algorithm; Carl Zeiss Meditec, Dublin, CA). Four test locations were chosen for testing, including at least two with significantly reduced sensitivity that remained $\geq$ 6 dB (i.e., not perimetrically blind), with the four locations spaced around the visual field in all four quadrants to promote stable fixation during testing. Frequency-of-seeing curves were measured at each location for both size III and size V stimuli using the method of constant stimuli[34] on an Octopus perimeter (Haag-Streit, Köniz, Switzerland) via the Open Perimetry Initiative interface.[35] For the size III stimulus, seven contrasts were selected for testing at 3-dB intervals centered at the perimetric sensitivity (the mean at that location over the last two clinical visual field tests). For the size V stimulus, the contrasts tested were 4 dB higher, because increasing the stimulus area is expected to increase sensitivity.[36] At the two most damaged locations of the four selected for a given eye, the highest contrast stimulus to be tested was always set to 3.7 dB, the greatest contrast presentable by the Octopus perimeter. For each stimulus size, 35 presentations were made per contrast level per location, split into five runs to reduce fatigue, with runs alternating between size III

and size V stimuli. All protocols were approved and monitored by the Legacy Health Institutional Review Board and adhered to the Health Insurance Portability and Accountability Act of 1996 and the tenets of the Declaration of Helsinki. All participants provided written informed consent when all of the risks and benefits of participation had been explained to them.

For each location and each stimulus size, the proportion seen at each contrast was calculated. A cumulative Gaussian curve was fit to the frequency-of-seeing data, where $Proportion = FP + (Max - FP) \times \Phi[(Contrast - Mid)/SD]$. Here, FP represents the false-positive rate, as measured from 50 blank presentations interspersed within the runs; Φ represents a cumulative Gaussian distribution function, such that $\Phi(-\infty) = 0$, $\Phi(0) = 0.5$, and $\Phi(\infty) = 1$; and $Contrast$ represents the stimulus contrast for that presentation. The remaining three parameters are fit by constrained maximum likelihood estimation. Mid represents the midpoint of the curve and is constrained to be ≥ -10 dB (to ensure algorithmic convergence); SD represents the standard deviation of the curve, constrained to be ≥ 0 dB; and Max represents the maximum response rate to an arbitrarily high contrast stimulus. Conventionally, this would be assumed to equal 100% minus the false-negative rate, but we have previously shown that this asymptotic maximum can be well below 100% at damaged locations.[29,30] From this equation, the perimetric contrast sensitivity was calculated using the conventional definition in clinical perimetry—namely, the contrast giving 50% response probability (this would exactly equal Mid if and only if $Max = 100\% - FP$). All analyses were performed using the statistical programming language R 4.0.3 (R Foundation for Statistical Computing, Vienna, Austria).

Next, for each location and stimulus size, 500 simulated frequency-of-seeing curves were generated, with individual stimulus responses simulated by randomly sampling from a binomial distribution with probability equal to the observed response probability at that contrast. The same fitting procedure was used to determine the simulated sensitivity and standard deviation. Simulated curves with sensitivity > 40 dB, sensitivity < -10 dB, or standard deviation > 40 dB were omitted. The intra-location variances were then calculated for sensitivity, SD, and log(SD), as a metric of the expected measurement error of each parameter.

The variability (SD) of each frequency-of-seeing curve was plotted against the perimetric contrast sensitivity for each stimulus size. Previously, Henson et al.[10] reported that variability (defined, as here, as the standard deviation of a cumulative Gaussian fit to frequency-of-seeing data but assuming that the upper and lower asymptotes of the fit were 100% and 0%,

respectively) for a size III stimulus increased exponentially with glaucomatous damage. They found a best fit model of the form $\log_e(SD) = -0.081 \times$ sensitivity $+ 3.27$, using locations with sensitivities between approximately 10 and 37 dB. Since that paper was published,[10] it has become apparent that characteristics of perimetric sensitivity estimates may alter when sensitivity declines to approximately 28 dB, because Ricco's area expands in glaucoma.[26] At ~28 dB, it can become larger than the size III stimulus,[37] so that the response characteristics are influenced by complete spatial summation rather than (until that point) incomplete summation.[27,38,39] It is not yet clear whether this affects variability. Thus, we repeated their exponential model fitting of the form $\log_e(SD) = A \times$ sensitivity $+ B$, excluding locations for which the sensitivity was above 28 dB. We also excluded locations with sensitivity estimates below 3.7 dB (i.e., based on the curve being extrapolated beyond the highest contrast presented). For comparison, we also fit a linear model of the form $SD = A_{Lin} \times$ sensitivity $+ B_{Lin}$. In each case, the regression lines were determined using Deming regression to account for measurement errors in both sensitivity and SD,[40,41] using the ratio between the squared measurement errors derived from the simulations described in the previous paragraph. In both cases, variability was compared between size III and size V, for matched sensitivity, within the same range of 3.7 to 28 dB, using a generalized estimating equation (GEE) model to account for intra-eye correlations.[42]

Experiment 2. Measuring Test–Retest Variability With Minimized Short-Term Variability

The second part of the study used repeated testing with a customized Zippy Estimation by Sequential Testing (ZEST) algorithm,[43,44] which tests just four visual field locations, with 30 presentations per location. The advantages of this over using clinical perimetry data when investigating long-term variability are that the short-term variability is as low as can reasonably be achieved within a short test duration (around 5 minutes per eye, similar to clinical perimetry, to avoid excessive fatigue effects) and is predictable using the results from Experiment 1. A size V stimulus was used for this experiment to further decrease the short-term variability.[32]

The 137 participants were recruited from the ongoing longitudinal Portland Progression Project.[45,46] For inclusion, they were required to have a diagnosis of glaucoma or suspected glaucoma in at least one eye, as determined by their clinician. Eyes with non-glaucomatous visual field loss were excluded. Each participant was tested twice, with a six-month interval between tests (or as close as their visit could be scheduled). One eye was tested per participant. In order to obtain a greater spread of sensitivity estimates, the most damaged eye was chosen for testing, except if two or more of the four chosen locations had sensitivity < 0 dB on their most recent visual field test, in which case the better eye was tested. Eyes that underwent any ocular surgery during that interval were excluded.

For the right eye, the four locations tested were $(9°, -15°)$, $(-15°, -9°)$, $(-9°, 15°)$, and $(15°, 9°)$; the locations were mirrored for the left eye. Thus, all four locations had the same mid-peripheral eccentricity, so that the slope of partial spatial summation is approximately equal between locations[39]; with one per quadrant to increase spatial uncertainty.[47] Testing was performed on an Octopus perimeter as in Experiment 1, with background intensity of 10 cd/m^2 and stimulus duration 200 ms, to match clinical perimetry as closely as possible. At each location, the Bayesian ZEST algorithm started by assuming a flat prior for sensitivity within the range of 10 to 45 dB. The lower end of this range was chosen because we have previously shown that sensitivities below 15 to 19 dB are unreliable, with variability obscuring any actual signal.[29] The prior distribution for sensitivities extended beyond the range of plausible true sensitivities in both directions to aid algorithmic convergence. After each presentation, the posterior PDF was calculated by multiplying the prior by either a cumulative Gaussian (if the participant responded to the stimulus) or its inverse (if not), with SD according to the equation of Henson et al.[10] The next stimulus contrast to be presented was chosen as the mean of the current prior distribution.[43] For each location, the sensitivity estimates after 10, 20, and 30 presentations were recorded.

The absolute error, $abs[(Sens_1 - Sens_2)/2)]$, was plotted against the estimated sensitivity, $(Sens_1 + Sens_2)/2$. Locally estimated scatterplot smoothing (LOESS) curve fitting[48] was used to predict the expected squared error for a given sensitivity, which equals the test–retest variance that would be predicted for a new participant with that sensitivity; the predicted SD (the square root of this variance) was added to the plot. Note that these LOESS fits were performed using variances, because variances of independent components of variability add linearly; however, results are presented as SDs (the square root of the variances) for easier visualization and interpretation.

Experiment 3. Calculating Long-Term Variability

Using results from Experiment 1, we could now create realistic frequency-of-seeing curves with different sensitivities. We repeatedly simulated the testing algorithm from Experiment 2, using these frequency-of-seeing curves, to estimate the short-term variability that would be expected if the true underlying sensitivity remained constant. By comparing this against the actual test–retest variability that was observed in Experiment 2, we could calculate the long-term variability in the true sensitivity, which would remain independent of the actual algorithm used.

For a chosen "true" sensitivity (*Sens*), a predicted frequency-of-seeing curve was generated using one of three models to derive its SD:

- Exponential model—$\log_e(\text{SD}) = A \times \text{sensitivity} + B$, using the best fit exponential model for sensitivities between 3.7 and 28 dB for a size V stimulus from Experiment 1.
- Henson model—$\log_e(\text{SD}) = -0.081 \times \text{sensitivity} + 3.27$, using the results of Henson et al.[10]
- Linear model—$\text{SD} = A_{Lin} \times \text{sensitivity} + B_{Lin}$, using the best fit linear model for sensitivities between 3.7 and 28 dB from Experiment 1. This model additionally assumed that the SD reached a floor when sensitivity > 28 dB (i.e., when the stimulus may be smaller than Ricco's area, causing partial spatial summation[37]); above this point SD was kept constant.

The ZEST algorithm from Experiment 2 was then simulated twice using each predicted frequency-of-seeing curve. If a simulated test–retest pair is denoted by $Sens_1$ and $Sens_2$, then this represents two "measurements" with an estimated sensitivity of ($Sens_1$ + $Sens_2$)/2 and an estimated squared error of [($Sens_1$ – $Sens_2$)/2]2. The absolute difference between the two sensitivity estimates after 10, 20, and 30 presentations can then be compared against the observed data from Experiment 2 for the same estimated sensitivity.

To match the distribution of sensitivities from Experiment 2, 100 test–retest pairs were simulated, with the "true" sensitivity being set to equal each of the participant's sensitivity estimates from Experiment 2 (after 30 presentations, then averaged between the two test dates). Hence, a total of 137 participants × 4 locations × 100 pairs = 54,800 pairs of simulated runs were generated for each model of the short-term variability. As in Experiment 2, the absolute errors were plotted against estimated sensitivity for all 54,800 simulated pairs, and a LOESS curve[48] based on the

squared errors was fit to the data. This LOESS fit provides the predicted mean squared error (i.e., the predicted test–retest variance from short-term variability alone) for the chosen model at any given sensitivity. The variance of the long-term variability was then calculated as the observed test–retest variance (from Experiment 2) minus this predicted short-term variance.

Results

Experiment 1. Characterizing Short-Term Variability

Each of the 35 participants (mean age, 69.9 years; range, 52–87) was tested at four locations of one eye. The mean deviation of the tested eye on the most recent clinic visit averaged −10.7 dB (range, −20.9 to −3.4). The perimetric sensitivities of the locations tested (i.e., the mean of the pointwise sensitivities at their last two clinic visits) averaged 18.9 dB (range, 4 to 32). Nineteen out of 140 tested locations had perimetric sensitivity ≤ 10 dB.

Figure 1 shows plots of variability (the standard deviation of the fitted frequency-of-seeing curve, or *SD*) against sensitivity (the contrast at which the participants would be predicted to respond to 50% of stimulus presentations based on the fitted frequency-of-seeing curve) for all tested locations for size III stimuli (Fig. 1A) and size V stimuli (Fig. 1B). The mean sensitivity of the tested locations was 18.6 dB for size III stimuli, and it was 25.4 dB for size V stimuli. The mean SDs were 3.6 dB for size III stimuli and 2.5 dB for size V stimuli. The expected measurement errors for size V stimuli were ±1.05 dB for sensitivity, ±1.24 dB for SD, and ±0.24 dB for log(SD). The solid red curves and the blue lines show the exponential and linear models, respectively, for locations with sensitivity ≥ 3.7 dB (so that sensitivity is not based on extrapolation beyond the range of contrasts tested) and ≤ 28 dB (so that the stimulus is smaller than Ricco's area[37]), using Deming regression based on the ratio of expected measurement errors. These lines were given by

Size III: $\log_e(\text{SD}) = 2.576 - 0.068 \times \text{sensitivity}$
$\text{SD} = 9.658 - 0.287 \times \text{sensitivity}$
Size V: $\log_e(\text{SD}) = 2.578 - 0.060 \times \text{sensitivity}$
$\text{SD} = 11.316 - 0.329 \times \text{sensitivity}$

At the same location, the size V stimulus resulted in higher sensitivity than the size III stimulus (mean difference = 9.6 dB; $P < 0.001$ from GEE regression) and lower variability (SD = 1.03 dB lower;
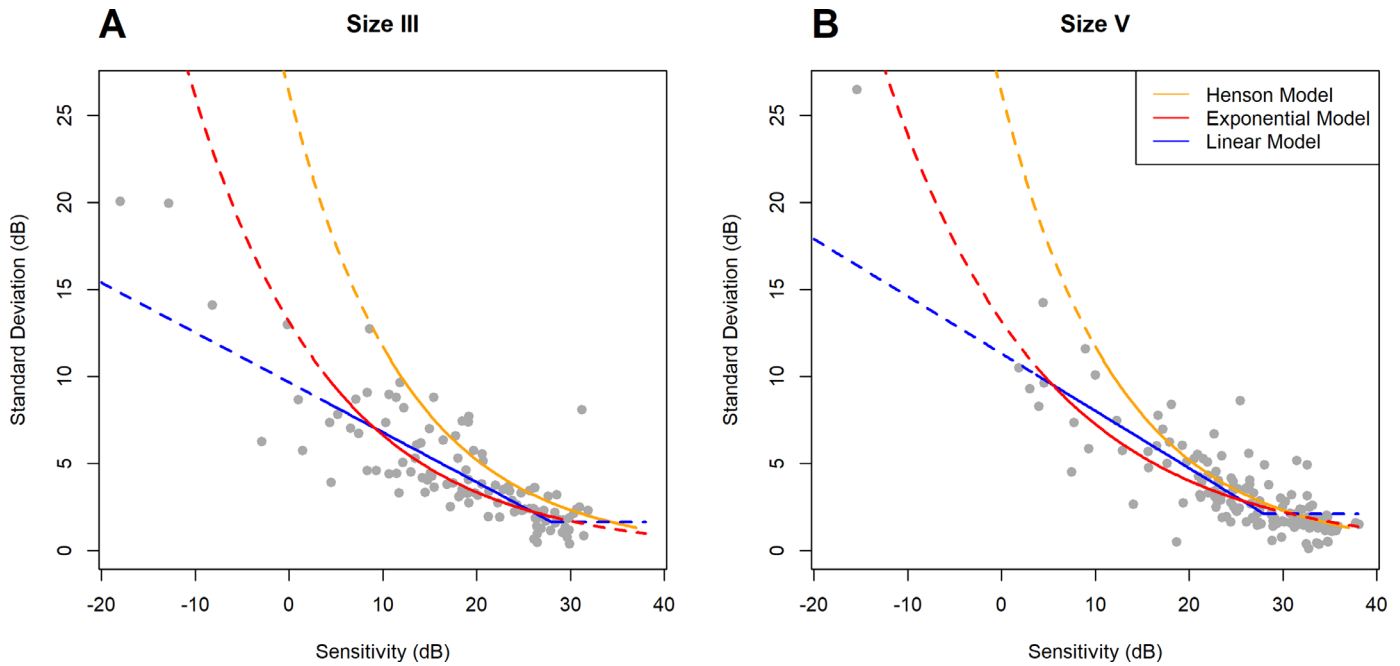
Figure 1. The relation between variability and sensitivity for two perimetric stimulus sizes based on frequency-of-seeing curves. Variability is defined as the standard deviation of a cumulative Gaussian fit to the response probabilities; sensitivity is defined based on the contrast that gives 50% response probability from the same fit. Each *gray symbol* represents one visual field location. The *solid curves* and *lines* show fits to the data for locations with sensitivity ranging from 3.7 to 28 dB for the exponential (*red*) and linear (*blue*) models (i.e., detection threshold was within the range of stimuli presented, and complete spatial summation can be assumed) or ≥ 10 dB for the Henson (*orange*) model (the range of measurements on which their model was based). Exponential and linear fits are based on Deming regression, with the ratio of measurement errors determined by the intra-location variance of estimates from 500 simulated frequency-of-seeing curves per location. The dashed lines extrapolate the fits beyond that range. For the linear fit, the extrapolated variability is set to be constant above 28 dB. (A) Using a size III stimulus (diameter, 0.43°). (B) Using a size V stimulus (diameter, 1.72°).

$P = 0.002$). However, for the same sensitivity, variability was slightly higher for the size V stimulus, whether using the linear model (SD = 0.75 dB higher; $P = 0.011$) or the exponential model, where log(SD) = 0.178 higher ($P = 0.027$). This difference did not depend on sensitivity ($P = 0.810$ for the linear model, $P = 0.247$ for the exponential model). The similarities between the fits for the two stimulus sizes in Figure 1, in particular the fact that the slopes do not significantly differ, are an indication that estimates of long-term variability derived by subtracting simulated short-term variability from observed test–retest variability should be the same for both sizes, as long as the simulation uses the appropriate model for that stimulus size. Thus, testing for Experiment 2 was conducted using the size V stimulus in order to further reduce short-term variability.

The linear model of short-term variability (blue line in Fig. 1) fit the data slightly better than the exponential model (red curve) within the range of 3.7 to 28 dB. For the size III stimulus, the root mean square error was 1.66 dB for the linear model versus 1.80 dB for the exponential model, and for the size V stimulus it was 1.69 dB for the linear model versus 1.71 dB for

the exponential model. A disadvantage of the linear model is that, if extrapolated to higher sensitivities above 35 dB, it would predict negative standard deviations. Thus, when using the linear model in simulations, we assumed that variability remains constant above 28 dB, an assumption that we do not have sufficient data to test. However, when the models were extrapolated to lower sensitivities as indicated by the dashed lines, the exponential model predicted unrealistically high variability. The formula of Henson et al.,[10] indicated by the orange curve, was based on locations with sensitivity ≥ 10 dB, and this higher floor caused it to overestimate variability at lower sensitivities (red dashed line) by even more. Notably, when locations with sensitivities > 28 dB were included in the fit (i.e., including locations at which complete spatial summation does not occur), the exponential model actually performed better than the linear model (root mean square error = 1.43 dB vs. 1.62 dB when using the size V stimulus). Thus, the recent and ongoing work examining the role of Ricco's area in perimetric sensitivities[26,27,37–39] may lead to re-evaluation of the optimal model of the sensitivity–variability relation.

Experiment 2. Measuring Test–Retest Variability With Minimized Short-Term Variability

The 137 participants were tested twice using the customized ZEST algorithm at four locations of one eye with 6 months between tests. The average age of participants was 72.3 years (range, 50–93). The average mean deviation of the tested eye on the date of the second visit, using the 24-2 visual field and SITA Standard testing algorithm,[33] was −1.79 dB (median, −0.92 dB; range, −19.85 to +2.87). For the four locations tested, the average SITA Standard sensitivity on the date of the second visit was 27.7 dB (range, <0 to +34), and 88 locations (17%) had sensitivity ≤25 dB. Using the custom-written, low-variability algorithm, the estimated sensitivities after 30 presentations per location were on average 0.70 dB higher ($P = 0.046$, GEE regression).

Figure 2 shows the estimated absolute error, abs($Sens_1 - Sens_2$)/2, plotted against estimated sensitivity, ($Sens_1 + Sens_2$)/2, after 10, 20, and 30 presentations. As expected, this test–retest variability increased with damage until sensitivity began to approach the measurement floor. The black line shows the LOESS fit to the data (noting that this LOESS fit is based on the squared errors, as explained above, so it is the square root of the predicted variance that is plotted as the black line). The apparent increase in variability at near-normal sensitivities when using 10 presentations per location (Fig. 2A) is due to the ZEST algorithm starting with a flat prior within the range of 10 dB to 45 dB, such that the first presentation was always 22.5 dB, and it took several presentations for the estimate to converge to values near 30 dB. The apparent increase in variability for sensitivities > 32 dB is due to very few locations having true sensitivities this high; hence, if the estimated sensitivity is >32 dB, then it is likely due to one of the two values, $Sens_1$ or $Sens_2$, being particularly noisy.

Experiment 3. Calculating Long-Term Variability

Figure 3 shows the predicted standard deviation of the short-term variability at different sensitivities based on a LOESS fit to simulated data; with the LOESS fit to the observed data from Experiment 2 (see Fig. 2) also shown for comparison. The distribution of "true" sensitivities used in the simulations was defined by the distribution of sensitivities after 30 presentations from Experiment 2, averaged between each participant's two visits. The exact same algorithm was used as in Experiment 2, with simulated responses based on random sampling with the probability of response to a given stimulus taken from a frequency-of-seeing curve with SDs as predicted using the linear (blue), exponential (red), or Henson (orange) model from Figure 1. As before, the LOESS fit was based on using squared errors to predict the variance before taking the square root of this prediction to derive the standard deviation shown in Figure 3.

Figure 4 shows the resulting estimates of long-term variability at different sensitivities. These are calculated as the observed test–retest variance (using the LOESS plot in Fig. 2) minus the predicted short-term variance (using the LOESS plot in Fig. 3). To aid interpretation, the *y*-axis in Figure 4 shows the SD of the long-term variability (i.e., square root of the calculated variance) rather than the variance itself. The estimates
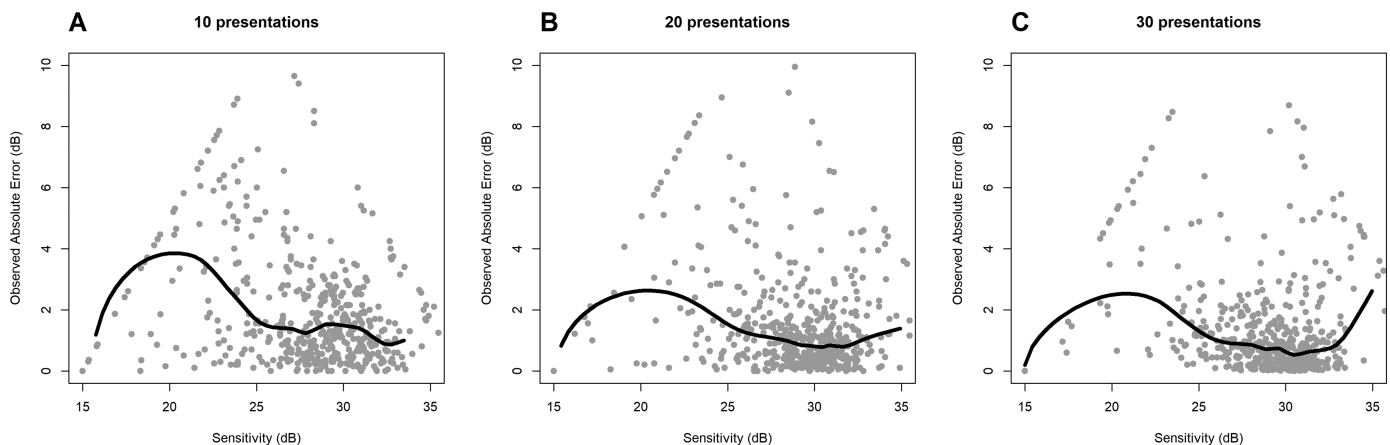

Figure 2. The estimated absolute error, abs($Sens_1 - Sens_2$)/2, plotted against estimated sensitivity, ($Sens_1 + Sens_2$)/2, after 10 (A), 20 (B), and 30 (C) presentations for four locations of 137 eyes tested twice 6 months apart. The *thick black curve* represents a LOESS fit to the data and indicates the predicted test–retest SD for a new participant with that sensitivity.
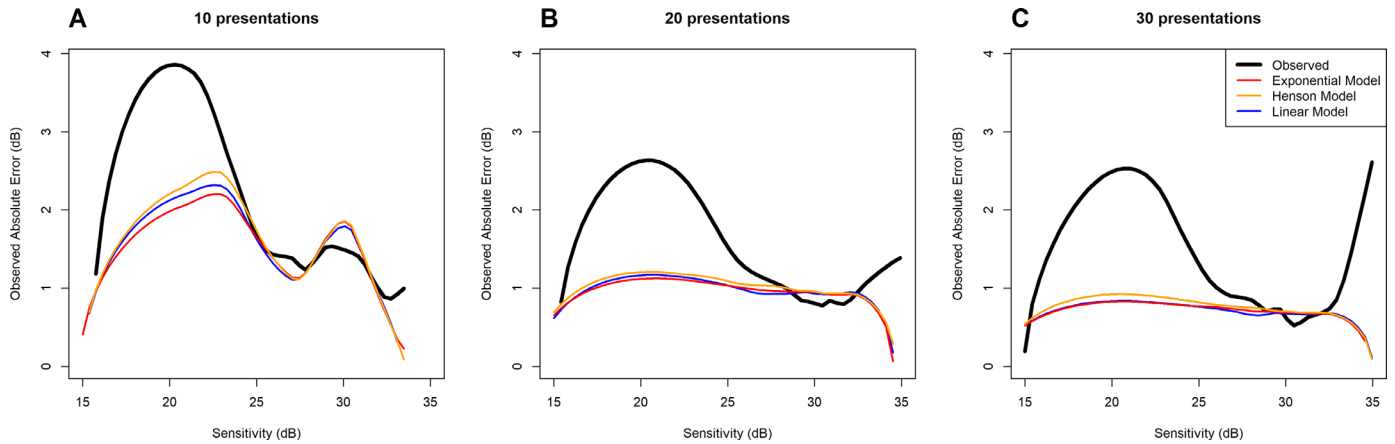
Figure 3. The predicted short-term variability of the testing algorithm used in Experiment 2 (expressed as a test–retest SD) for a given estimated sensitivity after 10 (A), 20 (B), and 30 (C) stimulus presentations. One hundred pairs of measurements were simulated for each location tested in Experiment 2 (i.e., 54,800 pairs in total). The "true" sensitivity for each simulation was set to equal the observed estimated sensitivity at the chosen location after 30 presentations, averaged between the two visits. The simulated frequency-of-seeing curve had SDs as predicted at the "true" sensitivity using the linear (*blue*), exponential (*red*), or Henson (*orange*) models from Experiment 1. LOESS fits were derived to predict the mean squared error, $[(Sens_1 - Sens_2)/2]2$, for a pair of simulated measurements with mean sensitivity $(Sens_1 + Sens_2)/2$; the square root of this prediction represents the expected SD. The *thick black curve* shows the equivalent LOESS fit derived from the observed test–retest data in Experiment 2 (as in Fig. 2).
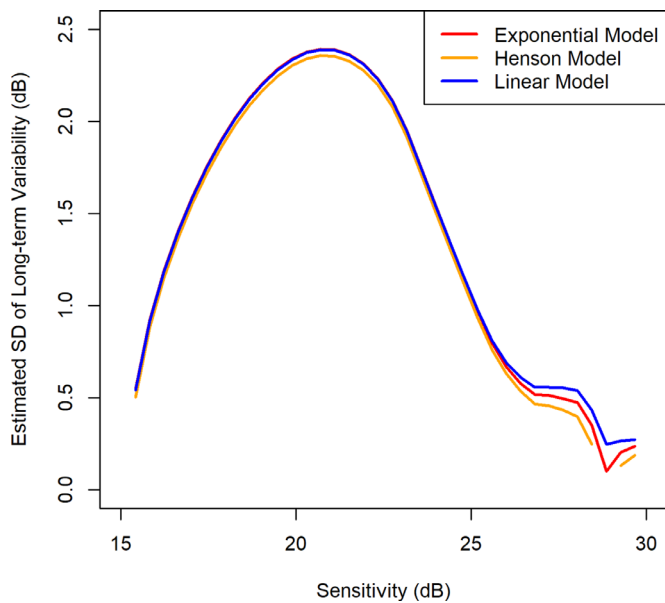

Figure 4. Estimated long-term variability for different sensitivities. Long-term variability was defined as the test–retest variance after 30 stimulus presentations from testing at a 6-month interval in Experiment 2 (see Fig. 2C) minus the short-term variance after 30 presentations from simulations in Experiment 3 (see Fig. 3C). It is displayed as a SD rather than variance to facilitate interpretation. The simulated short-term variability is based on the exponential (*red*), Henson (*orange*), or linear (*blue*) models from Experiment 1.

shown in Figure 4 are based on 30 presentations per location using the LOESS plots shown in Figures 2C and 3C. The estimates based on 20 presentations per location were almost identical. For the linear model,

the highest long-term variability estimate based on 30 presentations was 2.39 dB when the sensitivity was 20.7 dB; based on 20 presentations, the maximum was 2.36 dB at a sensitivity of 20.3 dB. Estimates based on 10 presentations were also very similar, but noisier. The estimates of long-term variability were also almost identical between the three models used for the simulations, further supporting the robustness of the results. As seen in Figure 3, the estimated observed test–retest variability (as fit by the LOESS curve) was actually smaller than the simulated short-term variability for sensitivities > 30 dB, an indication that long-term variability constitutes a negligible proportion of the test–retest variability at such locations.

Discussion

The primary aim of this study was to quantify the long-term variability of automated perimetry. By our definition and experimental design, this represents fluctuation in the true sensitivity and hence (unlike short-term variability) is independent of the testing algorithm used. Various factors have been suggested that may cause this kind of fluctuation. However, factors such as seasonality,[20] diurnal cycles, and technician experience[21] would seem to be independent of disease status; yet, we have demonstrated here that this long-term variability increases markedly in regions of glaucomatous loss. For estimated sensitivities of

30 to 32 dB, the observed test–retest SD was actually slightly smaller than the SD of the simulated short-term variability, as can be seen in Figure 3. The simulated variability may have been overestimated at these sensitivities, in particular due to the choice to assume constant variability in the linear model above 28 dB (see Fig. 1). However, we can still infer that, for locations with near-normal sensitivity, long-term variability can be considered to be negligible and test–retest variability is driven almost entirely by the short-term variability caused by the probabilistic nature of the frequency-of-seeing curve. At locations with glaucomatous damage, long-term variability increased. It reached a peak at sensitivities around 20 dB, below which the floor effect of the testing algorithm likely dominated and obscured any further increase.

In these damaged regions of the visual field, detection operates under complete spatial summation, and a small change in stimulus area has a greater effect on sensitivity than at healthy locations.[27,37–39] Similarly, a small movement in fixation will alter the number of retinal ganglion cells that are stimulated, and this will have a greater effect on sensitivity when complete spatial summation occurs, compared to healthy locations where only partial spatial summation occurs. Compounding this effect is that the number of cells stimulated may vary more in damaged regions than healthy regions for the same magnitude of fixation shift due to the inhomogeneous nature of glaucomatous loss. It is possible that factors such as fatigue also have a greater influence on sensitivity in a system that is already stressed by ongoing disease processes.[22,23] Additionally, pathophysiologic effects such as reduced axonal transport[49] and altered autoregulation of blood flow[50,51] could cause individual retinal ganglion cells to only function intermittently, causing variations in the true sensitivity at those locations. Transient scotomas have been demonstrated in healthy subjects after IOP elevation,[52] and glaucomatous eyes may be more susceptible to such changes.[53]

There are also consequences for clinical diagnostics. It has been suggested that the standard deviation of the test–retest variability of automated perimetry would have to be reduced by at least 20% to detect functional change one visit sooner without changing the testing frequency.[7] It has been reported that the root mean square test–retest error using the SITA Standard algorithm at a location with sensitivity 20 dB is approximately 4 dB,[54] meaning that a 0.8-dB reduction in variability would be required. Yet, the long-term variability constitutes 2.39 dB of this total, as suggested by our results above (Fig. 4), and that portion cannot be reduced by altering the testing algorithm. The remaining 1.61 dB is short-term variability that does depend on the testing algorithm. Thus, a requirement of a 0.8-dB reduction in variability would necessitate a 50% reduction of short-term variability from algorithm-dependent sources.

Our results therefore emphasize that adjusting the testing algorithms to reduce short-term variability may not be enough by itself to meaningfully improve the repeatability of perimetry. It is important to also reduce long-term variability, which can be greater than short-term variability in regions of glaucomatous loss.[24] Possible approaches could try to shorten the test to reduce fatigue; indeed, this may be one reason why the observed increase in test–retest variability from using the SITA Fast testing algorithm instead of SITA Standard is not as large as might have been predicted.[55] Another approach could be to alter the test to make it more engaging and less stressful for the test subject, again with the aim of reducing the effect of fatigue.

Using smaller stimuli could improve the uniformity of healthy sensitivities across the visual field[37] and might increase the detectability of defects, especially centrally.[38] However, as seen in Experiment 1, reducing stimulus size increases the standard deviation of the frequency-of-seeing curve and hence increases the range, of contrasts over which the test subject is unsure whether or not they saw the stimulus, making the test more mentally taxing. There is an inherent conflict between the desire to use test stimuli smaller than Ricco's area to increase the short-term signal-to-noise ratio[56] versus the likelihood that this may make the test more fatiguing for the patient and hence increase long-term variability.

We can also infer from the results in Experiment 1 that the Henson model of short-term variability[10] used in some simulation studies[57–59] may be suboptimal. It has previously been suggested that the Henson model overestimates variability at low sensitivities and that, when including such locations, an exponential model with different coefficients giving lower estimates could be more accurate[60]; this is consistent with the results of fitting the exponential model to our data from Experiment 1. It has also been suggested that the Henson model could be used, but a maximum standard deviation of 6 dB should be imposed to better represent empirical estimates of variability at low sensitivities.[31] However those studies and the original study by Henson et al.[10] all assumed that the relation between sensitivity and variability should be consistent across the range, of sensitivities. Yet, recent studies have shown that properties of sensitivity estimates from perimetry differ between locations where the stimulus area is within versus outside Ricco's area.[26,27,37–39] We propose that the same may be true of the relation

between sensitivity and variability. Data from this study and from previous studies[31,61] are actually more consistent with a linear model. Such a model would assume that, when sensitivity is below around 28 dB, variability increases linearly as sensitivity decreases. Such a model gives a very similar fit to the exponential model from 10 to 28 dB but without the excessively high estimates at lower sensitivities. The apparently better fit of the exponential model is solely due to the influence of locations with sensitivities above this cutoff. The exact upper cutoff that is optimal and the form of the model that could be used above that sensitivity cannot be determined from our data. The linear model used in the simulations of Experiment 3 assumed constant variability at locations above 28 dB, but that assumption was made for simplicity in the absence of better information. The exact sensitivity at which complete spatial summation ends will inevitably vary between individuals, and 28 dB is an approximation based on previous studies[37,39] rather than a definitively optimized value. Above whatever cutoff is chosen, it is certainly plausible that variability could still be related to sensitivity among locations undergoing partial spatial summation, just with a shallower slope.[56] It should be noted that, when results were based on 30 stimulus presentations per location, the model used did not appreciably impact the estimated short-term variability of the testing algorithm or the consequent estimate of the long-term variability, as seen in Figures 3C and 4. For the purposes of this study, an inaccurate model delays but does not prevent convergence to the sensitivity estimate. Thus, although this issue is important for future studies and indeed our results from Experiment 1 can contribute to those studies, it does not affect the validity of our conclusions concerning long-term variability.

Another interesting finding from Experiment 1 is the marked similarity between the sensitivity–variability relations that were found using a size III stimulus (Fig. 1A) versus a size V stimulus (Fig. 1B). In both cases, variability increased approximately linearly with glaucomatous damage below ∼28 dB; the predicted SD at a given sensitivity was within 1 dB of the same value for the two stimuli, and the slope (i.e., increase in SD for a 1-dB worsening of sensitivity) did not differ significantly between the stimulus sizes. The most relevant implication for this study is that this similarity supported using size V stimuli instead of size III for Experiment 2. More generally, it supports the idea that the perimetric response is driven primarily by the number of remaining retinal ganglion cells that are stimulated. If the density of remaining ganglion cells is halved but the stimulus area is doubled,

then this should result in very nearly the same sensitivity and variability. This is consistent with the finding that, if stimulus area is modulated while contrast remains constant, the response variability does not vary with defect depth.[27] Similarly, although it has been shown that Ricco's area increases in glaucoma,[26] the sensitivity to a stimulus of area exactly equaling Ricco's area should remain the same. In our linear model from Experiment 1, the transition from constant short-term variability above the cutoff to variability increasing linearly with loss below the cutoff occurs at the same cutoff of 28 dB for both size III and size V stimuli, even though increasing stimulus size means that more locations now have sensitivity greater than this value.[30] Interestingly, a very similar relation between variability and sensitivity for size III and V stimuli, with hints of a possible change at around 28 dB, has been found for patients with Leber hereditary optic neuropathy,[62] another condition in which scotomas result from retinal ganglion cell loss.

The estimates of long-term variability derived in this study are based on two test dates 6 months apart. It has been suggested that variability may increase with the time between test dates,[24] but we are not able to test that idea with the current dataset. The study by Urata et al.[24] found that the variability among five weekly tests was lower than the variability among five annual tests. Over longer time periods, the rate of change will often not be constant due to, for example, treatment changes, so the SD of residuals from a linear trend may overestimate the underlying long-term variability. It therefore remains unclear whether long-term variability would increase continually with the intertest interval or plateau when tests are more than some interval apart.

Testing for Experiment 2 was performed as part of a longitudinal testing protocol (Portland Progression Project), consisting of a series of diagnostic tests including standard clinical perimetry with the SITA Standard testing algorithm, retinal blood flow measurements, and a series of optical coherence tomography scans. It is possible then that the participants were more fatigued than a typical clinical patient. Counteracting this, as willing participants in a longitudinal study they are generally more interested and hence more motivated to maintain attention during the test, especially because they have more extensive interaction with the experienced testing technician than would normally be the case in a busy clinic. Another feature of the cohort is that they had extensive perimetric experience, and it is reasonable to assume that the observed test–retest variability would be higher in inexperienced patients.[63,64]

In summary, we report substantial long-term variability in the underlying pointwise sensitivity that cannot be explained by short-term intratest variability alone. Among locations with sensitivity greater than approximately 28 dB, where the perimetric stimulus may be larger than Ricco's area, short-term variability appears to vary little with sensitivity, and long-term variability constitutes only a very small proportion of the total test–retest variability. Among more damaged locations, where the perimetric stimulus is smaller than Ricco's area and so complete spatial summation occurs, short-term variability and long-term variability both increase with the severity of functional loss.

Acknowledgments

Disclosure: **S.K. Gardiner**, None; **W.H. Swanson**, None; **S.L. Mansberger**, None

References

1. Caprioli J. The importance of rates in glaucoma. *Am J Ophthalmol*. 2008;145(2):191–202.
2. Saunders LJ, Russell RA, Kirwan JF, McNaught AI, Crabb DP. Examining visual field loss in patients in glaucoma clinics during their predicted remaining lifetime. *Invest Ophthalmol Vis Sci*. 2014;55(1):102–109.
3. Heijl A, Lindgren A, Lindgren G. Test-retest variability in glaucomatous visual fields. *Am J Ophthalmol*. 1989;108(2):130–135.
4. Piltz JR, Starita RJ. Test-retest variability in glaucomatous visual fields. *Am J Ophthalmol*. 1990;109(1):109–111.
5. Garway-Heath DF, Lascaratos G, Bunce C, et al. The United Kingdom Glaucoma Treatment Study: a multicenter, randomized, placebo-controlled clinical trial: design and methodology. *Ophthalmology*. 2013;120(1):68–76.
6. Demirel S, De Moraes CG, Gardiner SK, et al. The rate of visual field change in the ocular hypertension treatment study. *Invest Ophthalmol Vis Sci*. 2012;53(1):224–227.
7. Turpin A, McKendrick AM. What reduction in standard automated perimetry variability would improve the detection of visual field progression? *Invest Ophthalmol Vis Sci*. 2011;52(6):3237–3245.
8. Blumenthal EZ, Sample PA, Zangwill L, Lee AC, Kono Y, Weinreb RN. Comparison of long-term variability for standard and short-wavelength automated perimetry in stable glaucoma patients. *Am J Ophthalmol*. 2000;129(3):309–313.
9. Hutchings N, Wild JM, Hussey MK, Flanagan JG, Trope GE. The long-term fluctuation of the visual field in stable glaucoma. *Invest Ophthalmol Vis Sci*. 2000;41(11):3429–3436.
10. Henson DB, Chaudry S, Artes PH, Faragher EB, Ansons A. Response variability in the visual field: comparison of optic neuritis, glaucoma, ocular hypertension, and normal eyes. *Invest Ophthalmol Vis Sci*. 2000;41(2):417–421.
11. Spry PG, Johnson CA, McKendrick AM, Turpin A. Variability components of standard automated perimetry and frequency-doubling technology perimetry. *Invest Ophthalmol Vis Sci*. 2001;42(6):1404–1410.
12. Anderson D, Patella V. *Automated Static Perimetry*. 2nd ed. St. Louis, MO: Mosby; 1999:147–159.
13. de Ruyter van Steveninck RR, Lewen GD, Strong SP, Koberle R, Bialek W. Reproducibility and variability in neural spike trains. *Science*. 1997;275(5307):1805–1808.
14. Warzecha A-K, Egelhaaf M. Variability in spike trains during constant and dynamic stimulation. *Science*. 1999;283(5409):1927–1930.
15. Reich DS, Victor JD, Knight BW, Ozaki T, Kaplan E. Response variability and timing precision of neuronal spike trains in vivo. *J Neurophysiol*. 1997;77(5):2836–2841.
16. Gardiner SK, Swanson WH, Demirel S, McKendrick AM, Turpin A, Johnson CA. A two-stage neural spiking model of visual contrast detection in perimetry. *Vision Res*. 2008;48(18):1859–1869.
17. Klein SA. Measuring, estimating, and understanding the psychometric function: a commentary. *Percept Psychophys*. 2001;63(8):1421–1455.
18. Strasburger H. Converting between measures of slope of the psychometric function. *Percept Psychophys*. 2001;63(8):1348–1355.
19. Flammer J, Drance S, 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(5):704–706.

20. Gardiner SK, Demirel S, Gordon MO, Kass MA, Ocular Hypertension Treatment Study Group. Seasonal changes in visual field sensitivity and intraocular pressure in the ocular hypertension treatment study. *Ophthalmology*. 2013;120(4):724–730.

21. Junoy Montolio FG, Wesselink C, Gordijn M, Jansonius NM. Factors that influence standard automated perimetry test results in glaucoma: test reliability, technician experience, time of day, and season. *Invest Ophthalmol Vis Sci*. 2012;53(11):7010–7017.

22. Johnson CA, Adams CW, Lewis RA. Fatigue effects in automated perimetry. *Appl Opt*. 1988;27(6):1030–1037.

23. Hudson C, Wild JM, O'Neill EC. Fatigue effects during a single session of automated static threshold perimetry. *Invest Ophthalmol Vis Sci*. 1994;35(1):268–280.

24. Urata CN, Mariottoni EB, Jammal AA, et al. Comparison of short- and long-term variability in standard perimetry and spectral domain optical coherence tomography in glaucoma. *Am J Ophthalmol*. 2020;210:19–25.

25. Khuu SK, Kalloniatis M. Standard automated perimetry: determining spatial summation and its effect on contrast sensitivity across the visual field. *Invest Ophthalmol Vis Sci*. 2015;56(6):3565–3576.

26. Redmond T, Garway-Heath DF, Zlatkova MB, Anderson RS. Sensitivity loss in early glaucoma can be mapped to an enlargement of the area of complete spatial summation. *Invest Ophthalmol Vis Sci*. 2010;51(12):6540–6548.

27. Rountree L, Mulholland PJ, Anderson RS, Garway-Heath DF, Morgan JE, Redmond T. Optimising the glaucoma signal/noise ratio by mapping changes in spatial summation with area-modulated perimetric stimuli. *Sci Rep*. 2018;8:2172.

28. Pan F, Swanson WH, Dul MW. Evaluation of a two-stage neural model of glaucomatous defect: an approach to reduce test-retest variability. *Optom Vis Sci*. 2006;83(7):499–511.

29. Gardiner SK, Swanson WH, Goren D, Mansberger SL, Demirel S. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage. *Ophthalmology*. 2014;121(7):1359–1369.

30. Gardiner SK, Demirel S, Goren D, Mansberger SL, Swanson WH. The effect of stimulus size on the reliable stimulus range of perimetry. *Transl Vis Sci Technol*. 2015;4(2):10.

31. Rubinstein NJ, Turpin A, Denniss J, McKendrick AM. Effects of criterion bias on perimetric sensitivity and response variability in glaucoma. *Transl Vis Sci Technol*. 2021;10(1):18.

32. Wall M, Kutzko KE, Chauhan BC. Variability in patients with glaucomatous visual field damage is reduced using size V stimuli. *Invest Ophthalmol Vis Sci*. 1997;38(2):426–435.

33. Bengtsson B, Olsson J, Heijl A, Rootzen H. A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmol Scand*. 1997;75(4):368–375.

34. Laming D, Laming J. F. Hegelmaier: on memory for the length of a line. *Psychol Res*. 1992;54(4):233–239.

35. Turpin A, Artes PH, McKendrick AM. The Open Perimetry Interface: an enabling tool for clinical visual psychophysics. *J Vis*. 2012;12(11):22.

36. Choplin NT, Sherwood MB, Spaeth GL. The effect of stimulus size on the measured threshold values in automated perimetry. *Ophthalmology*. 1990;97(3):371–374.

37. Phu J, Khuu SK, Zangerl B, Kalloniatis M. A comparison of Goldmann III, V and spatially equated test stimuli in visual field testing: the importance of complete and partial spatial summation. *Ophthalmic Physiol Opt*. 2017;37(2):160–176.

38. Choi AYJ, Nivison-Smith L, Khuu SK, Kalloniatis M. Determining spatial summation and its effect on contrast sensitivity across the central 20 degrees of visual field. *PLoS One*. 2016;11(7):e0158263.

39. Gardiner SK. Differences in the relation between perimetric sensitivity and variability between locations across the visual field. *Invest Ophthalmol Vis Sci*. 2018;59(8):3667–3674.

40. Deming W. *Statistical Adjustment of Data*. New York: Wiley; 1943.

41. Marín-Franch I, Malik R, Crabb DP, Swanson WH. Choice of statistical method influences apparent association between structure and function in glaucoma. *Invest Ophthalmol Vis Sci*. 2013;54(6):4189–4196.

42. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.

43. King-Smith PE, Grigsby SS, Vingrys AJ, Benes SC, Supowit A. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Res*. 1994;34(7):885–912.

44. Turpin A, McKendrick AM, Johnson CA, Vingrys AJ. Properties of perimetric threshold estimates from full threshold, ZEST, and SITA-like strategies, as determined by computer simulation. *Invest Ophthalmol Vis Sci*. 2003;44(11):4787.

45. Gardiner SK, Johnson CA, Demirel S. Factors predicting the rate of functional progression in early and suspected glaucoma. *Invest Ophthalmol Vis Sci*. 2012;53(7):3598–3604.

46. Gardiner SK, Mansberger SL, Demirel S. Detection of functional change using cluster trend analysis in glaucoma. *Invest Ophthalmol Vis Sci*. 2017;58(6):BIO180–BIO90.

47. Pelli DG. Uncertainty explains many aspects of visual contrast detection and discrimination. *J Opt Soc Am A*. 1985;2(9):1508–1532.

48. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*. 1979;74(368):829–836.

49. Fahy ET, Chrysostomou V, Crowston JG. Minireview: impaired axonal transport and glaucoma. *Curr Eye Res*. 2016;41(3):273–283.

50. Pournaras CJ, Rungger-Brandle E, Riva CE, Hardarson SH, Stefansson E. Regulation of retinal blood flow in health and disease. *Prog Retin Eye Res*. 2008;27(3):284–330.

51. Wang L, Cull G, Burgoyne CF, Thompson S, Fortune B. Longitudinal alterations in the dynamic autoregulation of optic nerve head blood flow revealed in experimental glaucoma. *Invest Ophthalmol Vis Sci*. 2014;55(6):3509–3516.

52. Drance SM. Studies in the susceptibility of the eye to raised intraocular pressure. *Arch Ophthalmol*. 1962;68:478–485.

53. Henkind P. Symposium on glaucoma: joint meeting with the national society for the prevention of blindness: new observations on the radial peripapillary capillaries. *Invest Ophthalmol Vis Sci*. 1967;6(2):103–108.

54. Artes PH, Iwase A, Ohno Y, Kitazawa Y, Chauhan BC. Properties of perimetric threshold estimates from Full Threshold, SITA Standard, and SITA Fast strategies. *Invest Ophthalmol Vis Sci*. 2002;43(8):2654–269.

55. Saunders LJ, Russell RA, Crabb DP. Measurement precision in a series of visual fields acquired by the standard and fast versions of the Swedish interactive thresholding algorithm: analysis of large-scale data from clinics. *JAMA Ophthalmol*. 2015;133(1):74–80.

56. Rountree L, Mulholland PJ, Anderson RS, Morgan JE, Garway-Heath D, Redmond T. Quantifying the signal/noise ratio with perimetric stimuli optimised to probe changing spatial summation in glaucoma. *Invest Ophthalmol Vis Sci*. 2017;58(8):2852.

57. Gardiner SK, Crabb DP. Examination of different pointwise linear regression methods for determining visual field progression. *Invest Ophthalmol Vis Sci*. 2002;43(5):1400–1407.

58. Chong LX, McKendrick AM, Ganeshrao SB, Turpin A. Customized, automated stimulus location choice for assessment of visual field defects. *Invest Ophthalmol Vis Sci*. 2014;55(5):3265–3274.

59. Denniss J, McKendrick AM, Turpin A. Towards patient-tailored perimetry: automated perimetry can be improved by seeding procedures with patient-specific structural information. *Transl Vis Sci Technol*. 2013;2(4):3.

60. Russell RA, Crabb DP, Malik R, Garway-Heath DF. The relationship between variability and sensitivity in large-scale longitudinal visual field data. *Invest Ophthalmol Vis Sci*. 2012;53(10):5985–5990.

61. Chauhan BC, Tompkins JD, LeBlanc RP, McCormick TA. Characteristics of frequency-of-seeing curves in normal subjects, patients with suspected glaucoma, and patients with glaucoma. *Invest Ophthalmol Vis Sci*. 1993;34(13):3534–3540.

62. Mejia-Vergara AJ, Sadun AA, Chen AF, Smith MF, Wall M, Karanjia R. Benefit of stimulus size V perimetry for patients with a dense central scotoma from Leber's hereditary optic neuropathy. *Transl Vis Sci Technol*. 2021;10(12):31.

63. Heijl A, Bengtsson B. The effect of perimetric experience in patients with glaucoma. *Arch Ophthalmol*. 1996;114(1):19–22.

64. Gardiner SK, Demirel S, Johnson CA. Is there evidence for continued learning over multiple years in perimetry? *Optom Vis Sci*. 2008;85(11):1043–1048.