Article

Choice of Stimulus Range and Size Can Reduce Test-Retest Variability in Glaucomatous Visual Field Defects

William H. Swanson¹, Douglas G. Horner¹, Mitchell W. Dul², and Victor E. Malinovsky¹

¹ Indiana University School of Optometry, Bloomington, IN, USA

² SUNY College of Optometry, New York, NY, USA

Correspondence: William H. Swanson, Indiana University School of Optometry, 800 East Atwater Avenue, Bloomington, IN 47405, USA. e-mail: wilswans@indiana.edu

Received: 4 April 2014

Accepted: 20 July 2014

Published: 25 September 2014

Keywords: automated perimetry, psychophysics, variability

Citation: Swanson WH, Horner DG, Dul MW, Malinovsky VE. Choice of stimulus range and size can reduce test-retest variability in glaucomatous visual field defects. *Trans Vis Sci Tech.* 2014;3(5):6, <http://tvstjournal.org/doi/full/10.1167/tvst.3.5.6>, doi: 10.1167/tvst.3.5.6

Purpose: To develop guidelines for engineering perimetric stimuli to reduce test-retest variability in glaucomatous defects.

Methods: Perimetric testing was performed on one eye for 62 patients with glaucoma and 41 age-similar controls on size III and frequency-doubling perimetry and three custom tests with Gaussian blob and Gabor sinusoid stimuli. Stimulus range was controlled by values for ceiling (maximum sensitivity) and floor (minimum sensitivity). Bland-Altman analysis was used to derive 95% limits of agreement on test and retest, and bootstrap analysis was used to test the hypotheses about peak variability.

Results: Limits of agreement for the three custom stimuli were similar in width (0.72 to 0.79 log units) and peak variability (0.22 to 0.29 log units) for a stimulus range of 1.7 log units. The width of the limits of agreement for size III decreased from 1.78 to 1.37 to 0.99 log units for stimulus ranges of 3.9, 2.7, and 1.7 log units, respectively ($F = 3.23$, $P < 0.001$); peak variability was 0.99, 0.54, and 0.34 log units, respectively ($P < 0.01$). For a stimulus range of 1.3 log units, limits of agreement were narrowest with Gabor and widest with size III stimuli, and peak variability was lower ($P < 0.01$) with Gabor (0.18 log units) and frequency-doubling perimetry (0.24 log units) than with size III stimuli (0.38 log units).

Conclusions: Test-retest variability in glaucomatous visual field defects was substantially reduced by engineering the stimuli.

Translational Relevance: The guidelines should allow developers to choose from a wide range of stimuli.

Introduction

Test-retest variability of visual field results obtained by conventional static automated perimetry is greater within and near glaucomatous visual defects when compared to regions with normal sensitivity.^{1–7} As a consequence, visual fields must be repeated several times in order to reliably determine whether individual patients are stable or progressing. Increased variability also adds time and complexity to clinical trials that attempt to assess effects of treatment.^{8–10} Increased test-retest variability is not present for frequency-doubling perimetry, which uses large, sinusoidal stimuli.^{11–15} Knowledge about why frequency-doubling stimuli have this property could help identify a wider range of potential new stimuli for perimetry with improved test-retest variability near and within glaucomatous defects.

The use of sinusoidal stimuli in perimetry has been referred to¹⁶ as “contrast sensitivity” perimetry. Perimetry with frequency-doubling stimuli is a form of contrast sensitivity perimetry that uses the large, hard-edged sinusoidal gratings that were common in basic vision science three decades ago. As display technology advanced, hard-edge sinusoids were replaced in basic research with soft-edged stimuli, most often Gaussian-windowed sinusoids,¹⁷ which are often referred to as “Gabor” stimuli. Perimetry with Gabor stimuli is similar to perimetry with frequency-doubling stimuli in that test-retest variability does not increase in glaucomatous defects.¹⁸ These variability properties of sinusoidal stimuli could be due to increased stimulus size^{19–21} or to decreased stimulus range.^{22–24} The current study was designed to assess the effects of stimulus size and range and to provide

guidelines for selecting perimetric stimuli that reduce test-retest variability.

Sinusoidal stimuli are larger than the size III stimulus used in conventional perimetry. The size III stimulus was defined by Goldman²⁵ and is a sharp-edged circular luminance increment with a width of 0.4°. By comparison, frequency-doubling stimuli are sharp-edged square stimuli with widths of 10°, 5°, and 2°. ^{12,13,26} It is possible that the size of sinusoidal stimuli, rather than their property of being sinusoidal, is the source of reduced variability. For instance, variability in glaucomatous defects can also be reduced¹⁵ by using the size V stimulus, which has a width of 1.7°. However, size alone may not be sufficient to reduce variability because even size V and frequency-doubling stimuli have been found to have an effective dynamic range that is smaller than the range of possible stimulus contrasts provided by the instrument.²⁷

Sinusoidal stimuli usually have a maximum contrast near 100% because they contain luminance decrements whose amplitude cannot exceed the mean luminance. Stimuli such as size III and size V are simple luminance increments without a corresponding decrement, so the maximum is only limited by the device. For conventional perimetry the maximum contrast for size III and size V is greater than 10,000% and is only limited by the choices of the manufacturer. However, high contrasts produce physiological saturation of retinal ganglion cell responses.^{22,28} Saturation is a basic property of all neurons that generate action potentials: Firing rate has a maximum set by factors such as refractory periods of ion channels, and responses become nonlinear when cells are strongly stimulated. Once stimulus contrast is high enough that ganglion cell responses become nonlinear, then further increases in stimulus contrast will cause minimal increases in the neuronal signal.²⁹ Recordings from primate ganglion cells²² have found that responses to the size III stimulus become nonlinear by 300% contrast (20 dB in clinical perimetry), and ganglion cell saturation has been predicted to cause high variability when high contrasts are used.^{23,24} If this is the case, then it is the limited range of stimulus contrasts of sinusoids that accounts for reduced variability for frequency-doubling stimuli and not the fact that they are sinusoids.

The current study assessed these factors. Experiment 1 compared variability for sinusoidal and nonsinusoidal stimuli using the same sizes and stimulus ranges, and experiment 2 assessed the effects of stimulus range on variability for conventional size

III perimetry. Results of these two experiments yielded guidelines for choosing custom stimuli that reduce variability. The effectiveness of the guidelines was examined in experiment 3, which compared long-term variability for perimetry with size III, frequency doubling, and custom stimuli.

Methods

Subjects

Sixty-two patients with glaucoma and 41 age-similar control subjects free of eye disease were tested as part of a multicenter longitudinal study³⁰ in three cities: Bloomington and Indianapolis, Indiana (Indiana University School of Optometry) and New York City, New York (SUNY College of Optometry).

Experiment 1 tested 20 glaucoma patients ages 46 to 84 years (mean 62.8, SD 10.4) and 10 age-similar control subjects ages 47 to 77 years (mean 57.2, SD 8.1). All were experienced with perimetry using clinical systems and had participated in prior studies with custom testing stations presenting custom stimuli. The time between visits ranged from 2 to 28 days (mean 11.5, SD 8.5).

Experiments 2 and 3 analyzed data from 51 patients with glaucoma and 37 controls; 9 patients and 6 controls had also participated in experiment 1. The age range was 47 to 85 years (mean 68.6, SD 8.5) for the patients with glaucoma and 47 to 77 years (mean 61.9, SD 8.3) for the controls. The time between visits ranged from 10 to 35 weeks (mean 17, SD 6).

For each patient, the diagnosis of glaucoma was made by the treating clinician based on a comprehensive ophthalmic examination including medical history, refraction, best-corrected visual acuity, slit lamp biomicroscopy (including gonioscopy), applanation tonometry, dilated funduscopy, stereoscopic ophthalmoscopy of the optic disc with a 78-diopter (D) lens, stereo photos of the optic nerve, and optic nerve imaging. A wide range of visual field damage was represented: Mean deviation (MD) for size III perimetry ranged from -18.4 to +1.3 dB (mean -4.9, SD 5.0 dB), and pattern standard deviation (PSD) ranged from 1.6 to 15.7 dB (mean 6.3, SD 4.1 dB).

For each control subject, one of the clinicians participating in the study reviewed the results of a recent comprehensive eye examination and data in the study chart to ensure that the person was found to be free of eye disease. MD for size III perimetry ranged

from -2.5 to $+1.4$ dB (mean 0.0 , SD 0.9 dB), and PSD ranged from 0.9 to 2.1 dB (mean 1.5 , SD 0.3 dB).

Inclusion criteria for both groups were as follows: age 45 to 85 years at time of enrollment, absence of known eye disease during a comprehensive eye examination within 2 years (except for glaucoma in the patient group), best corrected visual acuity of 20/20 or better (20/30 over age 70), spherical equivalent within -6 D to $+2$ D, cylinder correction within 3 D, clear ocular media. Exclusion criteria were an ocular or systemic disease (except for glaucoma in the patient group) known to affect the visual field (e.g., diabetic retinopathy, prior vein occlusion, macular degeneration, degenerative myopia, migraines), history of intraocular surgery (except uncomplicated cataract surgery more than a year before enrollment or past glaucoma surgery in the patient group), use of medications known to affect vision, inability to produce consistent data on perimetry, and pupil diameter less than 3 mm. Additional exclusion criteria for patients with glaucoma were intraocular pressure (IOP) > 30 mm Hg on a recent clinic visit and failure to attend routine clinic visits. Additional exclusion criteria for controls were IOP > 21 mm Hg, a first-degree relative with glaucoma, narrow angles and/or peripheral anterior synechiae, abnormal optic disc appearance (definite signs consistent with glaucoma, such as regional rim narrowing or notching or wedge-shaped retinal nerve fiber layer defects), and abnormal fundus appearance.

Institutional review board (IRB) approval was obtained from IRBs at Indiana University and SUNY College of Optometry. Written informed consent was obtained from each participant after explanation of the procedures and goals of the study before testing began.

Equipment

For the Gabor and Gaussian blob stimuli, testing was performed using custom testing stations based on cathode-ray tube (CRT) displays driven by a visual stimulus generator (ViSaGe; Cambridge Research Systems, Ltd., Cambridge, UK) that provided a resolution of 800×600 pixels with 14-bit control of luminance of each pixel. A photometer with calibration software (Opti-Cal; Cambridge Research Systems, Ltd.) was used to measure luminance versus voltage values for each phosphor, calculate transfer functions, and produce red, green, and blue (RGB) gamma correction look-up tables. The calibrations were checked periodically during the study, and they have remained stable. These experiments used 21-inch

monitors: Radius PressView 21SR (Miro Displays, Inc., Braunschweig, Germany) with a frame rate of 152 Hz at SUNY and Diamond Pro 2070SB (Mitsubishi Digital, Irvine, CA) with a frame rate of 140 Hz at Indiana University.

A custom-built motorized headrest controlled head position, and a 30-cm test distance was used so that the display subtended $55^\circ \times 46^\circ$ of visual angle, allowing testing of the peripheral locations of the central visual field as tested by the 24-2 test pattern. Two adjustable headrests allowed compensation for prominence of the brow, and custom 50-mm spherical lenses were held in place by magnets so that the metal rim remained as a cue to head and eye position (a plain glass lens was used when no spherical correction was needed). The patient's head was placed in an X-Y motorized chinrest and positioned so that the pupil was centered in the corrective lens (checked with a webcam) that was centered on the fixation target at a distance of 30 cm. Subjects were corrected for refractive error (spherical equivalent) and the 30-cm test distance. The appropriateness of the correction was checked by measuring acuity in the apparatus at the 30-cm test distance before testing. The smallest stimulus used on these systems was a Gaussian blob with 0.25° SD, or about four pixels. This means that approximately 50 pixels represent ± 1 SD, and approximately 100 pixels represent ± 2 SD. The Gaussian blob has a simple spatial form (e.g., no stripes, monotonic decline), so 50 to 100 pixels are sufficient to render it. If there were pixel artifacts, then sensitivity would be affected by blur, but these stimuli have been found to be resistant to effects of blur.³¹

For the size III stimulus, testing was performed on the Humphrey Field Analyzer (HFA) (Carl Zeiss Meditec, Inc., Dublin, CA), using the SITA-standard algorithm³² with its 24-2 pattern of stimulus locations shown in Figure 1. For the frequency-doubling stimulus, testing was performed with the Humphrey Matrix (Carl Zeiss Meditec, Inc.), using a ZEST algorithm^{33,34} and its 24-2 pattern of stimulus locations as shown in Figure 1.

Perimetry

Five forms of perimetry were used, two clinical tests and three custom³¹ tests. One form of custom stimulus was a simple luminance increment, known as a blob stimulus,³⁵ that is defined by the standard deviation of the Gaussian window. The other form of custom stimulus was a Gabor sine,³¹ a sine wave grating multiplied by a Gaussian window. Examples

Figure 1. Stimuli and stimulus locations. The *left panel* shows grayscale images for the full range of stimuli sizes used in the experiments. The *right panel* shows the three sets of stimulus locations used in the experiments

of the stimuli used are shown in the left panel of [Figure 1](#). The top row shows, from left to right, size III, 0.25° blob, 0.5° blob, size V, 0.50 cycle/deg Gabor. The size V stimulus was not used in these experiments and is shown for reference. The second row shows the largest blob (1.11°) on the left and a 0.25 cycle/deg Gabor on the right. The third row shows the largest Gabor (0.14 cycle/deg), and the fourth shows the frequency-doubling stimulus (0.5 cycle/deg in a 5° square window). Size III (top left) and frequency-doubling (bottom) stimuli are dramatically different in size: The frequency-doubling stimulus is 5° wide, and the size III stimulus is a circle less than one-tenth as wide. Both size III and frequency-doubling stimuli have hard edges, with the entire stimulus at full contrast from edge to edge. The custom stimuli span most of this range of sizes, but their soft edges make them larger than size III. The soft edges are formed by a two-dimensional Gaussian window that gives maximum contrast in the center of the stimulus and a smooth transition to the background luminance.

For two forms of custom perimetry, one with blobs and one with Gabor sines, the stimuli were magnified

based on retinal location, using a previously derived³⁶ spatial scale factor.³⁷ The third form of perimetry used 0.5° blobs of a fixed size ($SD = 0.5^\circ$) at all locations; these had been found to be very resistant to effects of blur.³¹ Both types of blob stimuli were presented at 55 conventional locations, the 54 size III 24-2 locations and a stimulus presented behind the fixation target. Frequency-doubling perimetry used its own set of 24-2 test locations. The Gabor sinusoids were presented at 57 custom locations,³¹ 13 identical to size III 24-2 locations and 44 biologically inspired.^{38,39} The right panel of [Figure 1](#) shows the three sets of test locations.

The temporal presentation for size III perimetry on the HFA is a 200-msec rectangular increment in luminance. This conventional temporal presentation was also used for perimetry with 0.5° blobs and scaled blobs. The temporal presentation for frequency-doubling perimetry is a 500-msec increment in contrast, with ramped onset and offset, using 18-Hz counterphase square-wave flicker and no change in mean luminance.¹³ The temporal presentation for Gabor stimuli was a 600-msec increment in contrast,

with abrupt onset and offset of 5-Hz counterphase square-wave flicker. A temporal frequency of 5 Hz was chosen to avoid effects of reduced retinal illumination that affect frequency-doubling perimetry,^{40,41} and abrupt onset and offset were used to maximize the temporal contrast.

Each Gabor patch was a horizontal grating in sine phase within a two-dimensional Gaussian window whose size was increased as spatial frequency decreased.^{31,36} This means that the stimuli were magnified using the following formula: Standard deviation times spatial frequency equals 0.25. The stimuli were in sine phase so that their mean luminance was the same as the background, 50 candela per square meter (cd/m^2). This is half of the 100 cd/m^2 background for frequency-doubling perimetry. The background luminance for perimetry with blobs was 10 cd/m^2 , the same as that used for size III perimetry. This allowed Weber contrasts for the blobs of up to 900%, which is referred to as 15 dB on the HFA.

The custom-testing station employed a ZEST algorithm³⁴ for all three custom tests, which terminated at a given location after six presentations. The interstimulus interval averaged 1200 msec, with a variable foreperiod. In order for data from a single visit to be included in the analysis, all three tests were required to have rates of false positive and fixation loss responses no greater than 20%. Further detail on strategies for assessing and managing fixation loss, false positives, false negatives, lens and lid, and other artifacts are described elsewhere.^{18,36}

Clinical perimeters use units of decibels, which has a different meaning for each device. In order to have common units across devices, we converted results from clinical devices to log contrast sensitivity (logCS), where contrast sensitivity is the reciprocal of contrast threshold. Contrast is a unitless measure defined as $(\text{peak} - \text{mean}) / \text{mean}$; for luminance increments, this is equivalent to Weber contrast, and for sinusoids this is equivalent to Michelson contrast.²⁶ For size III perimetry, the device reports decibels as $25 + 10 \times \log\text{CS}$. For frequency-doubling perimetry the device reports decibels as $20 \times \log\text{CS}$.

Protocol

One eye of each subject was tested on all visits. Experiment 1 used the custom-testing stations to perform three custom perimetric tests whose order was counterbalanced across visits and subjects. The first visit was considered practice, and the data were

not used. The data from the second and third visits were used to assess test-retest variability for the patients with glaucoma. The data from the controls' second visit were used to compute mean normal values so that patient data could be expressed as *defect depth*, computed as log difference from mean normal. The ZEST algorithm provided a stimulus range of 1.7 log units for all three tests, from 16% to 800% Weber contrast for the blobs and from 1.2% to 61.3% Michelson contrast for the Gabors.

Data collection for experiment 1 was initiated a year before data collection for experiments 2 and 3, and these results were used to select the custom stimulus used in a prospective study of longitudinal fluctuations for perimetry with size III, frequency-doubling, and Gabor stimuli. The data analyzed were from the two most recent visits in this concluding longitudinal study where reliability criteria were met for all three forms of perimetry: frequency-doubling and size III perimetry on commercial perimeters and the custom test that had lowest variability in experiment 3.

Statistical Design

For all three experiments, the effective stimulus range was varied by imposing a floor and a ceiling. Whenever a measured value was lower than the floor, it was set equal to the floor, and whenever a measured value was higher than the ceiling, it was set equal to the ceiling. Results from experiment 1 guided choice of stimulus ranges in experiment 2, and results from experiments 1 and 2 guided choice of stimulus range in experiment 3.

For each experiment, two methods were used to compare test-retest variability across different data-sets. The first method used Bland-Altman⁴² analysis for descriptive statistics. For each test, the difference in log sensitivities at the two visits was plotted as a function of the mean of the log sensitivities, with each data point from a given location and a given patient. The standard deviation of the test-retest differences across all locations and subjects gave an estimate of overall test-retest variability across all levels of sensitivity and was used to compute the width of the 95% limits of agreement ($2 \times 1.96 \times \text{SD}$). The emphasis for Bland-Altman analysis was on descriptive statistics because a Bonferroni correction for eight possible comparisons (three for experiment 1, two for experiment 2, and three for experiment 3) would require $P < 0.0062$, and with $N = 50$, this would require a reduction in standard deviation by approximately

43% for statistical significance; the few instances with more than a 43% reduction ($F > 2.06$) were reported as statistically significant.

The second method used a bootstrap strategy,⁴³ both for descriptive statistics and for hypothesis testing about peak test-retest variability. For each data point on the Bland-Altman plot, the log difference (y -axis) was converted to its absolute value for use as the magnitude of the test-retest difference, and the mean (x -axis) was used as log defect depth (experiments 1 and 3) or logCS (experiment 2) for the data point.

The bootstrap method generated repeated runs with a random number generator, drawing the same size sample of data points for each run, then grouped the data points into bins of equal size and computed mean sensitivity and mean magnitude of difference for each bin. Each run allowed the same data point to be drawn multiple times. For descriptive statistics, two-tailed 95% confidence limits were computed by repeating the random selection 200 times, using the mean of the 200 values for a bin, along with the fifth highest and fifth lowest of the 200 values. For hypothesis testing, one-tailed 99% confidence limits were computed using 1000 repetitions and the fifth highest and fifth lowest of the 1000 values. Fourteen bins were used for all reported analyses; varying bin number did not substantially change the results.

Experiment 2 assessed effects of stimulus range for only the size III stimulus, so bins could be grouped by logCS. Experiments 1 and 3 compared variability across different stimuli, so the sensitivities were converted to depth of defect before grouping in bins. For each location and stimulus, the mean logCS was computed for the control group, and then depth of defect for a patient was expressed as the log difference between the patient's logCS and the mean of controls.

There were five hypotheses about test-retest variability, tested with one-tailed values, $P < 0.01$ for statistical significance using a Bonferroni correction. For experiment 1, the hypothesis was that the scaled sinusoidal stimuli would give lower test-retest variability than the scaled Gaussian blob stimuli. For experiment 2, the hypothesis was that variability would decrease as stimulus range was decreased. For experiment 3, the hypothesis was that variability would be lower for the custom stimuli and for the frequency-doubling stimulus than for the size III stimulus. Each hypothesis was tested with both methods: F -test on standard deviation of difference

scores in Bland-Altman analysis and overlap of confidence intervals of the means in bootstrap analysis.

Results

Figure 2 shows data for experiment 1 in Bland-Altman scatter plots, with 1120 data points in the scaled Gabor (top) graph and 1060 data points in the other two graphs. Random noise by ± 0.02 log units has been added in x and y so that data points with identical values can be distinguished. The stimulus range was 1.7 log units, and the limits of agreement were similar for the Gabors, fixed blobs, and scaled blobs at 0.72, 0.73, and 0.79 log units, respectively.

Figure 3 shows results of the bootstrap analysis for experiment 1 for test-retest variability as a function of log defect. Symbols show the means for sensitivity and variability for each bin, and error bars show the 95% confidence intervals for these means.

With the 1.7 log unit stimulus range (top panel), the lowest bin had a deeper average defect for the blob stimuli than for the Gabor stimuli, so the analysis was repeated with a stimulus range of 1.3 log units (bottom panel). For all three tests, for both stimulus ranges, the mean test-retest value was approximately 0.1 log units near mean normal (defect depth = 0.0) and increased to peak variability for an average defect depth between -0.5 and -0.9 log units. Peak variability was 0.22 log units for the Gabors at both stimulus ranges and declined with stimulus range for the blob stimuli: from 0.29 to 0.23 log units for the scaled blobs and from 0.26 to 0.22 log units for the big blobs. This led to the choice of the 1.7 log unit stimulus range for experiment 2 and the 1.3 log unit range for experiment 3. Confidence limits for test-retest variability of the three tests overlapped for most bins, and for seven bins with defects from -0.1 to -0.5 log units the mean test-retest variability tended to be lower for the scaled Gabors. This led to the choice of the scaled Gabors in the longitudinal study, represented in experiment 3.

Figure 4 shows the data from experiment 2, perimetry with the size III stimulus for three different stimulus ranges, in the same format as Figure 2. The ceiling was set to 37 dB, and three floors were used: -2 dB (0 dB stimulus was not seen), 10 dB, and 20 dB. This resulted in stimulus ranges of 3.9, 2.7, and 1.7 log units, for which the width of the 95% limits of agreement were 1.78, 1.37, and 0.99 log units, respectively. The stimulus range of 3.9 log units yielded a significantly larger width for the limits of

agreement than did the stimulus range of 1.7 log units ($F = 3.18$, $P < 0.0001$).

Figure 5 shows results of bootstrap analysis for experiment 2. The lowest bin was near the limit imposed by the stimulus range for all three floors, as were the two lowest bins for the floor of -1.5 log units and the four lowest bins for the floor of -0.5 log units. For all three stimulus ranges, the mean test-retest value was approximately 0.15 log units near mean normal (defect depth = 0.0) and increased in defects to a peak variability of 0.99, 0.54, and 0.34 log units, for stimulus ranges of 3.9, 2.7, and 1.7 log units, respectively. Peak variability was significantly lower for each decrease in stimulus range (no overlap of 99% confidence limits, $P < 0.01$).

Figure 6 shows Bland-Altman scatter plots for experiment 3, with stimulus range 1.7 log units for all three tests by use of floors and ceilings. The width of the limits of agreement was smallest for Gabor stimuli and largest for size III. Therefore, a stimulus range of 1.3 log units was used for the bootstrap analysis, shown in Figure 7. For all three tests, the lowest two bins were near the limit imposed by the stimulus range, as was the highest bin. Peak variability was 0.38, 0.24, and 0.18 log units for size III, frequency-doubling, and Gabor stimuli, respectively. The 99% confidence limits for Gabor and frequency-doubling stimuli did not overlap the confidence limits for peak variability for size III ($P < 0.01$).

Discussion

This study developed guidelines for engineering stimuli to yield low test-retest variability and developed a custom perimetric test that was successful in providing low test-retest variability in glaucomatous defects. These results provide guidance for design of perimetric stimuli and confirm predictions of neural modeling of the pathophysiology of glaucoma.^{22,26}

Imaging measures can be helpful in assessing progression in patients with early stages of disease, in large part because of lower between-subject variability than found in perimetry.⁴⁴ However, once a defect has reached an average of -5 dB, imaging measures reach a floor. For instance, peripapillary retinal nerve fiber layer thickness for an optic disc

←
Figure 2. Bland-Altman plots for data from experiment 1. Gray diamonds show the limits imposed by the stimulus range of 1.7 log units. The 95% limits of agreement (dashed lines) are shown, and the legend gives their width (LoA) in log units.

Figure 3. Results of bootstrap analysis of the absolute values of test-retest differences for data in Figure 2. *Top panel* shows results for a floor of -1.5 log units, *bottom panel* shows results for a floor of -1.1 log units.

sector may be similar for an average perimetric defect of -5 dB or -30 dB.⁴⁴ At the same time, when perimetry with size III returns a defect of -5 dB at a visual field location on one visit, it may return a defect of -20 dB or 0 dB on the next visit.³ The ability to reduce test-retest variability in defects has the potential⁸ to make perimetry a more useful clinical tool in patients with severe damage, where imaging measures are no longer useful.

The most important factor in reducing variability with size III was controlling the stimulus range by reducing effects of high stimulus contrasts. Frequency-doubling stimuli have an inherent maximum contrast of 100%, while size III perimetry uses contrasts greater than 10,000% because luminance increments have no inherent maximum contrast. Reducing the maximum contrast to 300% (-0.5 log contrast sensitivity) caused a 2.9-fold reduction in peak test-retest variability. A contrast of 300% for size III on the HFA corresponds to 20 dB, which is the lower limit for size III derived by Wall and colleagues²⁷ and near the lower limits derived by Gardiner and colleagues.²⁴ The floor represents not seeing the maximum stimulus, and on the HFA it is assigned the value -2 dB. In Figure 4, the floor was reached at least once for test and retest with size III at 3.1% of locations for a floor of -2 dB (upper panel of Fig. 4), 5.5% of locations for a floor of 10 dB (middle panel), and 10.9% of locations for a floor of 20 dB (lower panel). By comparison, in Figure 6 the floor was reached at least once for 3.8% of locations with scaled Gabors and 4.4% of locations with frequency-doubling stimuli.

The use of sinusoidal stimuli did not seem to have much effect on test-retest variability, beyond the fact that choice of sinusoidal stimuli prevents the use of high contrasts. Experiment 1, which used a 1.7 log unit stimulus range to avoid high contrasts, found little difference in variability for Gaussian blobs and the Gabors, so there seems to be no inherent advantage in using sinusoids as long as the stimulus range is restricted. Experiment 3 used Gabors because they yielded the lowest variability in experiment 1, but this does not mean that Gabors are optimal stimuli. For instance, the 0.5° blob has greater resistance to effects of blur,³¹ and peak test-retest variability was

Figure 4. Bland-Altman plots for data from experiment 2 in the same format as Figure 2, but for size III perimetry with three different floors: -2.7 log units (-2 dB), -1.5 log units (10 dB), -0.5 log units (20 dB).

Figure 5. Results of bootstrap analysis of the absolute values of test-retest differences for data in Figure 4, with stimulus ranges of 3.9, 2.7, and 1.7 log units (39, 27, and 17 dB).

not greater than for Gabors when higher contrasts were avoided (Fig. 3, lower panel).

The effect of stimulus size on test-retest variability was more complex than the role of stimulus range in that the frequency-doubling stimuli were much larger than the fixed blobs, yet had similar peak variability: 0.24 vs. 0.22 log unit, respectively. Compared to the size III stimulus, the fixed blobs had soft edges and were somewhat larger (0.5° SD for blobs versus radius of 0.2° for size III). The soft edges and larger size for the blobs are important because primate ganglion cells are responsive when the size III stimulus is centered within $\pm 0.2^\circ$ of the midpoint of the ganglion receptive field and become unresponsive when the stimulus is 0.5° from the midpoint.²² This is in accord with a proposal that fixational instability in perimetry with $SD \sim 0.5^\circ$ can account for high variability for the size III stimulus in regions of the visual field with gradients in sensitivity^{21,45} or with “holes” in ganglion cell arrays damaged by glaucoma.^{20,26}

Both of these explanations for the effect are consistent with our finding that test-retest variability

may be decreased with only a modest increase in stimulus size when soft edges are used. The hard edge of the size III target means that a 0.5° displacement would stimulate responses from a completely different set of ganglion cells, so the soft edges of the blobs were designed to cause a smoother transition between responsive and not responsive ganglion cells. Furthermore, modeling of neural processes indicates that edges may be important for detection of Goldmann stimuli,⁴⁶ and modeling of geometrical optics indicates that peripheral defocus can reduce perimetric sensitivity when stimuli have sharp edges.³¹ Similarly, the size V stimulus is large enough to be resistant to effects of blur,⁴⁷ so with a suitable floor it may give results similar to the fixed blob. Further research is needed to determine whether smooth edges reduce variability for larger hard-edged stimuli such as the size V stimulus.

Both shape and size influence how a stimulus will be affected by forward light scatter, which increases the effective size of the stimulus on the retina while reducing retinal contrast. For luminance increments,

Swanson et al.

Figure 6. Bland-Altman plots for data in experiment 3 in the same format as Figures 2 and 4.

Figure 7. Results of bootstrap analysis of the absolute values of test-retest differences for data in Figure 6, with a stimulus range of -1.3 log units.

such as the Goldmann stimuli and the Gaussian blobs, increasing the size of the stimulus increases the total amount of light being added to the background, which increases the amount of scattered light that can stimulate more distant ganglion cells. The effects of light scatter will be more pronounced for the hard-edged Goldmann stimuli than for the Gaussian blobs because the blobs already have scatter built into them. For sinusoidal stimuli such as the frequency-doubling and Gabor stimuli, just as much light is subtracted from the background in the dark bars as is added to the background in the light bars, so the primary effect of light scatter will be to reduce the contrast between light and dark bars. For the hard-edged frequency-doubling stimulus, the high contrasts at the hard edges will result in light scatter from all four edges onto receptive fields of more distant ganglion cells. By comparison, for the Gabor stimuli the contrast peaks in the center and has the same soft edges as the Gaussian stimuli, so effects of light scatter on more distant ganglion cells should be less than for the frequency-doubling stimuli.

The primary analysis was on test-retest variability at individual locations, but clinically the MD is used as a summary for overall field loss, and a group of adjoining flagged locations is used to assess extent of visual field loss. Test-retest variability is much lower for MD than for individual locations,⁴⁸ as well as for perimetric sensitivities averaged in terms of optic nerve sectors that serve those retinal locations.⁴⁴ Figure 8 shows test-retest variability for both MD and sector averages for the subjects in experiment 3. This shows the same trend as in a prior study,⁴⁸ with variability slightly lower for larger stimuli. For MD, the limits of agreement spanned 0.55 log units for size III stimuli, 0.38 log units for frequency-doubling stimuli, and 0.29 log units for the Gabor stimuli. A floor of -1.0 log units reduced the span for the limits of agreement for size III to 0.48 log units, still significantly larger than the 0.29 log unit span for the Gabor stimuli ($F = 2.74$, $P < 0.00001$).

A classic study with the 30-2 test pattern and the full threshold algorithm reported that variability in patients with glaucoma was high in locations with

Figure 8. Bland-Altman plots for MD (*top panel*) and sector averages for inferior temporal (IT) optic disc sector, (*middle panel*) and superior temporal (ST) optic disc sector (*bottom panel*) for the subjects in experiment 3.

normal sensitivity when these were more peripheral locations.³ We used the 24-2 test locations, which include very little of their third zone, and repeated the bootstrap separately for their two inner zones using the data in experiment 3 with a floor of -1.1 log units (-11 dB). With size III, variability for the bin with mean -0.1 dB was 30% higher in the outer zone than the inner zone. For the Gabor stimuli, variability was only 4% high in the outer zone. These findings are consistent with both ganglion saturation²² and visual field gradients²¹ as explanations for elevated variability because, for size III, more central locations have higher mean normal sensitivities and more shallow visual field gradients while Gabors yield high mean normal sensitivities across the tested locations, with very slight gradients.

We followed the approach of this classic study³ by assessing variability as a function of defect depth in the bootstrap analysis. This approach assumes that the relations between variability and defect depth are the same in patients with mild, moderate, and severe

visual field loss. To test this assumption, we divided the patients into three groups based on MD for size III: MD worse than -6.1 dB, MD better than -1.8 dB, or MD between these two values. We repeated the bootstrap analysis for each group and found the same pattern for each group: the lowest test-retest variability was for Gabor and frequency-doubling stimuli, the highest variability was for size III, and when bins for different groups had similar mean defects, they also had similar test-retest variability.

The results of this study can be applied in several different ways by clinicians assessing progression. First, keep in mind that when perimetry with size III returns a defect of -5 dB at a location on one visit, it may return a defect of -20 dB or 0 dB on the next visit.³ For instance, when assessing progression at individual locations, be more concerned about a decline from 0 dB to -5 dB than a decline from -5 dB to -15 dB. Second, consider using larger stimuli in patients who have a number of test locations worse than -5 dB. Frequency-doubling and size V stimuli

may yield more reliable measurements than size III in patients with extensive visual field damage.¹⁵ Third, MD varies by up to ± 2.5 dB across much of its range, so do not be alarmed by seeing lower *P* values for MD in subsequent tests.

In summary, high test-retest variability in perimetry can be substantially reduced by restricting the stimulus range to avoid high contrasts and by modestly increasing stimulus size. The use of stimuli that reduce test-retest variability has the potential to improve the ability of clinicians to determine whether a patient is stable and to improve the ability of clinical trials to assess whether a treatment slows progression. The current study shows several examples of such stimuli and illustrates the principles for designing new forms of perimetry with acceptable test-retest variability.

Acknowledgments

Andrew Turpin, PhD (University of Melbourne Department of Computing and Information Systems), engineered the ZEST algorithm used in the custom-testing stations.

Financial support was provided by the National Institutes of Health (Bethesda, MD, USA) through Grants R01EY007716 and 5P30EY019008. The funding organization had no role in the design or conduct of this research.

This paper was accepted for presentation at the annual meeting of the Association for Research in Vision and Ophthalmology in Orlando, Florida, May 2014.

Disclosure: **W.H. Swanson**, None; **D.G. Horner**, None; **M.W. Dul**, None; **V.E. Malinovsky**, None

References

1. Flammer J, Drance SD, Zulauf M. Differential light threshold: short- and long-term fluctuation in patients with glaucoma, normal controls and patients with suspected glaucoma. *Arch Ophthalmol*. 1984;102:704–706.
2. Werner EB, Petrig B, Krupin T, Bishop KI. Variability of automated visual fields in clinically stable glaucoma patients. *Invest Ophthalmol Vis Sci*. 1989;30:1083–1089.
3. Heijl A, Lindgren A, Lindgren G. Test-retest variability in glaucomatous visual fields. *Am J Ophthalmol*. 1989;108:130–135.
4. Piltz JR, Starita RJ. Test-retest variability in glaucomatous visual fields [letter]. *Am J Ophthalmol*. 1990;109:109–111.
5. 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;41:417–421.
6. 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;43:2654–2659.
7. Russell RA, Malik R, Chauhan BC, Crabb DP, Garway-Heath DF. Improved estimates of visual field progression using Bayesian linear regression to integrate structural information in patients with ocular hypertension. *Invest Ophthalmol Vis Sci*. 2012;53:2760–2769.
8. Turpin A, McKendrick AM. What reduction in standard automated perimetry variability would improve the detection of visual field progression? *Invest Ophthalmol Vis Sci*. 2011;52:3237–3245.
9. Boland MV, Ervin AM, Friedman DS, et al. Comparative effectiveness of treatments for open-angle glaucoma: a systematic review for the U.S. Preventive Services Task Force. *Ann Intern Med*. 2013;158:271–279.
10. Sena DF, Lindsley K. Neuroprotection for treatment of glaucoma in adults. *Cochrane Database Syst Rev*. 2013;2:CD006539.
11. 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;40:648–656.
12. Spry PG, Johnson CA, McKendrick AM, Turpin A. Variability components of standard automated perimetry and frequency-doubling technology perimetry. *Invest Ophthalmol Vis Sci*. 2001;42:1404–1410.
13. Artes PH, Hutchison DM, Nicolela MT, LeBlanc RP, Chauhan BC. Threshold and variability properties of matrix frequency-doubling technology and standard automated perimetry in glaucoma. *Invest Ophthalmol Vis Sci*. 2005;46:2451–2457.
14. Horani A, Frenkel S, Blumenthal EZ. Test-retest variability in visual field testing using frequency doubling technology. *Eur J Ophthalmol*. 2007;17:203–207.

15. Wall M, Woodward KR, Doyle CK, Artes PH. Repeatability of automated perimetry: a comparison between standard automated perimetry with stimulus size III and V, matrix, and motion perimetry. *Invest Ophthalmol Vis Sci.* 2009;50:974–979.
16. Harwerth RS, Crawford ML, Frishman LJ, Viswanathan S, Smith EL 3rd, Carter-Dawson L. Visual field defects and neural losses from experimental glaucoma. *Prog Retin Eye Res.* 2002;21:91–125.
17. Cannon MW Jr., Fullenkamp SC. Perceived contrast and stimulus size: experiment and simulation. *Vision Res.* 1988;28:695–709.
18. Hot A, Dul MW, Swanson WH. Development and evaluation of a contrast sensitivity perimetry test for patients with glaucoma. *Invest Ophthalmol Vis Sci.* 2008;49:3049–3057.
19. 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;38:426–435.
20. 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;83:499–511.
21. Wyatt HJ, Dul MW, Swanson WH. Variability of visual field measurements is correlated with the gradient of visual sensitivity. *Vision Res.* 2007;47:925–936.
22. Swanson WH, Sun H, Lee BB, Cao D. Responses of primate retinal ganglion cells to perimetric stimuli. *Invest Ophthalmol Vis Sci.* 2011;52:764–771.
23. Swanson W, Sun H, Lee B, Cao D. Author response: frequency-doubling technology and parasol cells. *Invest Ophthalmol Vis Sci.* 2011;52:3759–3760.
24. Gardiner S, Swanson W, Goren D, Mansberger S, Demirel S. Assessment of the reliability of standard automated perimetry in regions of glaucomatous damage [published online ahead of print March 12, 2014]. *Ophthalmology.* 2014;121(7):1359–1369. doi: 10.1016/j.ophtha.2014.01.020.
25. Goldmann H. Fundamentals of exact perimetry. 1945. *Optom Vis Sci.* 1999;76:599–604.
26. Sun H, Dul MW, Swanson WH. Linearity can account for the similarity among conventional, frequency-doubling, and Gabor-based perimetric tests in the glaucomatous macula. *Optom Vis Sci.* 2006;83:455–465.
27. Wall M, Woodward KR, Doyle CK, Zamba G. The effective dynamic ranges of standard automated perimetry sizes III and V and motion and matrix perimetry. *Arch Ophthalmol.* 2010;128:570–576.
28. 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;48:1859–1869.
29. Kaplan E, Shapley RM. The primate retina contains two types of ganglion cells with high and low contrast sensitivity. *Proc Natl Acad Sci U S A.* 1986;83:2755–2757.
30. Swanson WH, Malinovsky VE, Dul MW, Malik R, Sutton B, Torbit J. Contrast sensitivity perimetry and clinical measures of glaucomatous damage. *Optom Vis Sci.* 2014. In press.
31. Horner DG, Dul MW, Swanson WH, Liu T, Tran I. Blur-resistant perimetric stimuli. *Optom Vis Sci.* 2013;90:466–474.
32. Bengtsson B, Olsson J, Heijl A, Rootzen H. A new generation of algorithms for computerized threshold perimetry, SITA. *Acta Ophthalmol Scand.* 1997;75:368–375.
33. 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;34:885–912.
34. 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;44:4787–4795.
35. Bijl P, Koenderink JJ, Toet A. Visibility of blobs with a Gaussian luminance profile. *Vision Res.* 1989;29:447–456.
36. Keltgen KM, Swanson WH. Estimation of spatial scale across the visual field using sinusoidal stimuli. *Invest Ophthalmol Vis Sci.* 2012;53:633–639.
37. Watson AB. Estimation of local spatial scale. *J Opt Soc Am A.* 1987;4:1579–1582.
38. Asaoka R, Russell RA, Malik R, Crabb DP, Garway-Heath DF. A novel distribution of visual field test points to improve the correlation between structure-function measurements. *Invest Ophthalmol Vis Sci.* 2012;53:8396–8404.
39. Jansonius NM, Schiefer J, Nevalainen J, Paetzold J, Schiefer U. A mathematical model for describing the retinal nerve fiber bundle trajectories in the human eye: average course, variability, and influence of refraction, optic disc size and optic disc position. *Exp Eye Res.* 2012;105:70–78.

40. Kogure S, Membrey WL, Fitzke FW, Tsukahara S. Effect of decreased retinal illumination on frequency doubling technology. *Jpn J Ophthalmol*. 2000;44:489–493.
41. Swanson WH, Dul MW, Fischer SE. Quantifying effects of retinal illuminance on frequency doubling perimetry. *Invest Ophthalmol Vis Sci*. 2005;46:235–240.
42. Bland JM, Altman DG. Measurement error. *BMJ*. 1996;313:744.
43. Gentle JE. *Computational Statistics*. Dordrecht, Netherlands: Springer; 2009.
44. Hood DC, Anderson SC, Wall M, Raza AS, Kardon RH. A test of a linear model of glaucomatous structure-function loss reveals sources of variability in retinal nerve fiber and visual field measurements. *Invest Ophthalmol Vis Sci*. 2009;50:4254–4266.
45. Wyatt HJ. Automated perimetry: using gaze-direction data to improve the estimate of scotoma edges. *Invest Ophthalmol Vis Sci*. 2011;52:5818–5823.
46. Pan F, Swanson WH. A cortical pooling model of spatial summation for perimetric stimuli. *J Vis*. 2006;6:1159–1171.
47. Anderson RS, McDowell DR, Ennis FA. Effect of localized defocus on detection thresholds for different sized targets in the fovea and periphery. *Acta Ophthalmol Scand*. 2001;79:60–63.
48. Wall M, Doyle CK, Zamba KD, Artes P, Johnson CA. The repeatability of mean defect with size III and size V standard automated perimetry. *Invest Ophthalmol Vis Sci*. 2013;54:1345–1351.